

# Making Things Happen

A THEORY OF CAUSAL EXPLANATION

---



JAMES WOODWARD

**MAKING THINGS HAPPEN**

OXFORD STUDIES IN PHILOSOPHY OF SCIENCE

*General Editor*

Paul Humphreys, University of Virginia

*Advisory Board*

Jeremy Butterfield

Peter Galison

Ian Hacking

Philip Kitcher

Richard Miller

James Woodward

*The Book of Evidence*

Peter Achinstein

*Science, Truth, and Democracy*

Philip Kitcher

*The Devil in the Details: Asymptotic Reasoning in Explanation,  
Reduction, and Emergence*

Robert Batterman

*Making Things Happen: A Theory of Causal Explanation*

James Woodward

# MAKING THINGS HAPPEN

A Theory of Causal Explanation

JAMES WOODWARD

**OXFORD**  
UNIVERSITY PRESS  
2003

# OXFORD

UNIVERSITY PRESS

Oxford New York

Auckland Bangkok Buenos Aires Cape Town Chennai

Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata

Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi

São Paulo Shanghai Taipei Tokyo Toronto

Copyright © 2003 by Oxford University Press, Inc.

Published by Oxford University Press, Inc.

198 Madison Avenue, New York, New York 10016

[www.oup.com](http://www.oup.com)

Oxford is a registered trademark of Oxford University Press

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of Oxford University Press.

Library of Congress Cataloging-in-Publication Data

Woodward, James, 1946–

Making things happen : a theory of causal explanation / James Woodward.  
p. cm.

Includes bibliographical references and index.

ISBN 0-19-51527-0; 0-19-518953-1 (pbk.)

1. Causation. 2. Explanation. 3. Science—Philosophy. I. Title.

BD541 W64 2003

122—dc21 2002192596

First published as an Oxford University Press paperback 2005

9 8 7 6 5 4 3 2 1

Printed in the United States of America  
on acid-free paper

## Acknowledgments

This book defends what I call a *manipulationist* or *interventionist* account of explanation and causation. According to this account, causal and explanatory relationships are relationships that are potentially exploitable for purposes of manipulation and control. They are also relationships that are *invariant* under interventions. These ideas have their roots in a long tradition of work in experimental design and econometrics that goes back to such writers as Fisher, Haavelmo, Frisch, Strotz, Wald, Simon, and Rubin and that continues, at least in some respects, in the work of Peter Spirtes, Clark Glymour, Richard Scheines, and Judea Pearl. I have tried to take these ideas out of the social scientific and biomedical contexts for which they were originally designed and show how they may be generalized to other areas of science. I have also tried to provide philosophical foundations for these ideas, drawing out their implications, comparing them with alternative approaches, and defending them from criticisms.

Writers in the grip of a single, overarching set of ideas sometimes tend to suppose that these ideas can be used to resolve all of the extant problems in their subject area. I fear that I have not been immune to this impulse. Even readers who are not entirely unsympathetic to my project may well conclude that I claim far too much for the manipulationist approach and that, though it may illuminate some aspects of cause and explanation, these notions are too complex and multifaceted to be fully captured by any single approach. This may be the right assessment at the end of the day, but I think there is considerable value in taking a single interrelated set of ideas, developing them systematically, and trying to push them as far as they will go. In this way, we can explore the resources of the approach and learn what it can and cannot do. At any event, this book is written in this spirit. Those who are drawn to more pluralistic approaches will, I hope, at least be persuaded that the manipulationist account illuminates an important strand in our notions of causation and explanation and one that has been relatively neglected.

I have benefited enormously from the advice, encouragement, comments, and criticisms of many friends and colleagues, both on this manuscript and in reaction to the talks and papers on which it is based. In this connection, I would particularly like to thank Nancy Cartwright, Fiona Cowie, David Freedman, Clark Glymour, Alan Hajek, Dan Hausman, Ned Hall, Chris Hitchcock, Paul Humphreys, Francis Longworth, Judea Pearl, and Jonathan Schaffer. I have

also benefited from general discussions about causation and explanation with Frank Arntzenius, Jim Bogen, Phil Dowe, Evelyn Fox Keller, Richard Healey, Kevin Hoover, Gurol Irzik, Philip Kitcher, Igal Kwart, Peter Machamer, Peter Menzies, Sandy Mitchell, David Papineau, Wes Salmon, Brian Skyrms, and Elliott Sober. The Division of Humanities and Social Science at Caltech has been a wonderfully supportive place to work, with stimulating colleagues in philosophy and in other disciplines. Work on this manuscript has also been supported by a grant from the National Science Foundation (SBR-9320097).

In addition to these acknowledgments, I owe debts of a different sort. A very substantial portion of this book was written at Children's Hospital, Los Angeles during 2000–2001, while my daughter Katie was being treated for an aggressive brain tumor. I am grateful beyond measure to this institution, to Dr. Martin Weiss, her surgeon, and to Dr. Douglas Hyder, her oncologist, for saving her life. I am also very grateful to the many friends and colleagues who provided help and support during this period and made it possible for me to continue to work.

Finally, I owe an incalculable debt to my wife, Julia, for her love and encouragement throughout the unconscionably long time it has taken me to finish this book. It is very literally true that it would not have been written but for her. For this, for believing in me, and for much else, I am grateful with all my heart.

# Contents

1. Introduction and Preview 3
  2. Causation and Manipulation 25
  3. Interventions, Agency, and Counterfactuals 94
  4. Causal Explanation: Background and Criticism 152
  5. A Counterfactual Theory of Causal Explanation 187
  6. Invariance 239
  7. Causal Interpretation in Structural Models 315
  8. The Causal Mechanical and Unificationist Models  
of Explanation 350
- Afterword 374
- Notes 376
- References 398
- Index 407



*This page intentionally left blank*

**MAKING THINGS HAPPEN**

*This page intentionally left blank*

## Introduction and Preview

### 1.1 The Centrality of Cause and Explanation

An interest in causes and explanations pervades our lives. We wonder why our cars won't start, why corn grows better in one field than another, why a friend seemed particularly happy or gloomy yesterday. Scientists wonder why elementary particles have the mass they do, why the universe is, at a sufficiently large scale, (nearly) flat with a uniform mass distribution, why there are so many noncoding regions in the human genome, and why the dinosaurs became extinct.

Given the centrality of these interests, it is not surprising that there have been many attempts to theorize about causation and explanation, both within and outside of philosophy. Philosophical concern with these topics dates back to (at least) Plato and Aristotle, as David Ruben (1990) reminds us. Claims about causation play a central role in the doctrines of virtually all philosophers in the early modern period, from Descartes and Locke to Hume and Kant. More recently, the development of the deductive-nomological (*DN*) model of explanation by writers like Carl Hempel and the elaboration of detailed alternatives to this model by writers like Wesley Salmon and Philip Kitcher have made explanation a central topic in philosophy of science. Outside of philosophy, one finds less self-conscious theorizing about explanation, but there are rich and extensive literatures in statistics, econometrics, cognitive psychology, and computer science on problems of causal inference and how best to understand the notion of causation.

Despite this intensive discussion, I think that it is fair to say that there is less consensus on the topics of explanation and causation in philosophy than there was three or four decades ago, when the *DN* model was widely accepted. Indeed, as elsewhere in philosophy, the past few decades of work on causation and explanation have been characterized by a proliferation of self-contained schools with surprisingly little mutual influence or communication. Consider, for example, the question of the role of counterfactuals in the characterization of causation and explanation. Although the counterfactual analyses of causation developed by David Lewis and his students (Lewis, 1973, [1979] 1986c, 2000) have been influential within some areas of philosophy, such as metaphysics, they have had relatively little impact on philosophy of science.

## 4 Making Things Happen

Moreover, the Lewisian tradition has ignored related work in statistics and econometrics that also draws on ideas about the connection between causation and counterfactuals (see, e.g., Pearl 2000a). Many philosophers of science have in turn dismissed treatments of causation and explanation that rely on counterfactuals as unclear or unscientific, despite the existence of a mathematically sophisticated literature outside of philosophy that takes just this form.

Within philosophy of science, we find writers who work on what they call probabilistic theories of causation (again, there is very little contact with relevant work in statistics and econometrics), writers who think of causation as involving the transmission of some physical quantity such as energy, and writers who propose to analyze the notion of causation in terms of the notion of a law of nature, again with relatively little cross-talk. Moreover, all this discussion has had surprisingly little impact on philosophers (e.g., those working in philosophy of psychology and biology) who are not themselves specialists in causation/explanation but who draw on ideas about these subjects in their own work. To the extent that there is any single dominant view among this group, it probably remains some hazy version of the *DN* model: it is assumed that even if Hempel didn't get all the details right, it is nonetheless correct that explanations and causal claims must in some way "involve" or "be backed" by "general laws." The questions of what a "law" is, or what "backed" means, or how the *DN* model could possibly be correct, given the widely accepted counterexamples to it (see chapter 4), are regarded as best left to specialists.

This book is an attempt to construct a comprehensive account of causal explanation that applies to a wide variety of causal and explanatory claims in different areas of science and in ordinary life, which engages some of the relevant literature from other disciplines and avoids at least some of the difficulties faced by previous philosophical treatments. This introductory chapter describes in a general way my approach to this subject and also sketches some of the major themes that are explored in subsequent chapters. One cannot provide a detailed defense of everything, even in a long book, and so part of my purpose in this chapter is also to lay bare some of the prejudices and preconceptions that I bring to this discussion.

As a number of writers have observed, the word "explain" is used in a wide variety of ways. One may explain the meaning of a word, how to solve a certain kind of differential equation, how to bake a cherry pie, what happened during an argument (where this is simply to describe in detail), or why one decided on a certain course of action (where this to offer an excuse or justification). The account of explanation that I present is not intended to capture any of these uses of "explain." Still less do I attempt to say what all possible uses of "explain" have in common. Instead, I focus on a much narrower range of explanatory activities—roughly, those that consist of providing *causal* explanations (in a broad sense of "causal," described in more detail below) of why some particular outcome or general pattern of outcomes occurs. A distinguishing feature of such explanations is that they show how what is explained depends on other, distinct factors, where the dependence in question

has to do with some relationship that holds as a matter of empirical fact, rather than for logical or conceptual reasons. Thus, for example, explanation in mathematics, if there is such a thing, is outside the scope of my discussion. My reason for proceeding in this way is that I believe that an account that attempts to capture the common elements in everything we may wish to call an explanation is unlikely to be very illuminating and unlikely to tell us much about what is distinctive about causal explanation and the role it plays in inquiry. In the theory of explanation, as in science itself, generality is not always a virtue.

This is by no means a judgment shared by all philosophers. For example, in an important paper, Paul Churchland (1989) adopts a much more expansive conception of explanation. For Churchland, virtually any activity that involves recognition, classification, or description amounts to “explanation.” A not unrelated view is implicitly adopted by philosophers who subscribe to the popular position that all inductive inference is a matter of “inference to the best explanation” and who hold that the only reason any hypothesis is ever adopted is that it explains some set of facts that serve as evidence for it. It is clear that if this view is to be remotely plausible, “explanation” must be understood in a very broad sense: we must be willing to say that when a naturalist describes a new species of ant on the basis of careful field observations, his description “explains” those observations and is adopted for that reason, or that when I count the number of sheep in a field and arrive at “twenty-five,” the result of my counting is (if correct) “explained” by the fact that there really are twenty-five sheep in the field, or that when I estimate the mean of some random variable in a population on the basis of a sample drawn randomly from the population, the population mean “explains” the sample mean, and so on.

As I suggest below, one reason for not adopting this inclusive notion of explanation is that it collapses or obscures an important distinction that scientists themselves make: the distinction between explanation and description. This distinction is a pervasive one in science. Biologists regularly distinguish between descriptive and classificatory activities, on the one hand, and explanation and the discovery of causal relationships on the other. In statistics and econometrics, we find a closely parallel distinction between “descriptive statistics,” including information about correlations, on the one hand, and information about causal and explanatory relationships on the other. Problems involving “inductive” inference from, say, correlations in samples to population correlations are seen as importantly different from problems of causal inference. One of my guiding assumptions in what follows is that an adequate theory of causation and explanation ought to make sense of such distinctions: it ought to make it clear how causal and explanatory information differs from mere description. Views that take all forms of classification and description to be explanatory fail to satisfy this constraint.

I said above that the notion of explanation in which I am interested is causal explanation, broadly conceived. As many readers will be aware, the role that causal information plays in explanation is itself a disputed question.

Some writers hold that all explanation (or at least all explanation of why some outcome occurs) must be causal, and other writers deny this, holding instead that there are noncausal forms of (why) explanation. Writers also differ about what counts as a "causal explanation." Thus, Wesley Salmon (1984) embraces a notion of causal explanation according to which this involves tracing spatiotemporally continuous causal processes and intersections of such processes and also holds that all genuine explanation must be causal in this sense. According to Salmon, an account that traces the subsequent motion of two billiard balls to their prior collision would count as a causal explanation, whereas a derivation of the equilibrium pressure of a gas from the ideal gas law and prior initial conditions would not count as explanatory, because it fails to trace individual causal processes. Graham Nerlich (1979), by contrast, is in rough agreement with Salmon about what counts as a causal explanation, but holds that there is an important noncausal form of explanation, which he calls geometrical explanation. He offers as an example the explanation of the trajectories of free particles in gravitational field by reference to the affine structure of space-time. Salmon would presumably deny that such appeals to space-time structure are explanatory.

Yet another distinction between causal and noncausal forms of explanation is due to Elliott Sober (1983); he contrasts explanations that trace the actual sequence of events leading up to some outcome, which he thinks of as causal, with what he calls *equilibrium explanations*, in which an outcome is explained by showing that a very large number of initial states of a system will evolve in such a way that it ends up in the outcome state that we wish to explain, but in which no attempt is made to trace the actual sequence of events leading up to that outcome. Thus, an explanation that traces the actual sequence of molecular collisions leading up to the current thermodynamic state of a gas, as characterized by such macroscopic variables as temperature and pressure, counts as a causal explanation, while a demonstration that almost all molecular configurations consistent with the initial temperature and pressure of the gas would result in its current macroscopic state would count as a noncausal equilibrium explanation.

Given this variety of characterizations of the notion of causal explanation, some stipulation or regimentation of usage is obviously necessary if our discussion is to go forward. As already intimated, I favor a broad notion of causal explanation according to which, roughly, any explanation that proceeds by showing how an outcome depends (where the dependence in question is not logical or conceptual) on other variables or factors counts as causal. I suggest below that the distinguishing feature of causal explanations, so conceived, is that they are explanations that furnish information that is potentially relevant to manipulation and control: they tell us how, if we were able to change the value of one or more variables, we could change the value of other variables. According to this conception, both derivations involving the ideal gas law and Sober's equilibrium explanations count as causal explanations. This "manipulationist" conception of causal explanation has the advantage of fitting a wide range of scientific contexts, especially in the social and behavioral sciences,

where investigators think of themselves as discovering causal relationships and constructing causal explanations, but where narrower notions of causal explanation, such as Salmon's, seem to be of very limited applicability (see chapter 8). It also has the advantage of exhibiting an intuitively appealing underlying rationale or goal for explanation and the discovery of causal relationships: if these are relationships that are potentially exploitable for purposes of manipulation, there is no mystery about why we should care about them.

## 1.2 What Should an Account of Causal Explanation Aim to Do?

My discussion so far has been framed in terms of providing an account or elucidation of causal and explanatory claims, and I speak below of interpreting such claims or capturing or clarifying their "content" or "meaning" in manipulationist terms. Although this is not a book about philosophical methodology, some brief discussion of how I conceive this project is in order. A common view is that there are just two possibilities: (1) one may be in the business of attempting to provide a "conceptual analysis" of "cause" and related locutions, where this is a matter of describing ordinary usage or scientific usage; alternatively, (2) one may engage in what Dowe (2000) calls "empirical analysis," where this is "to discover what causation is in the objective world" (p. 1).

My project is certainly different in its results from the kind of empirical analysis executed by Dowe. (I leave it to the reader to decide whether it counts as discovering "what causation is.") But although a significant portion of what I attempt does involve a description of ordinary and scientific usage and judgment, my project goes well beyond this—it is not just "conceptual analysis" in the sense described above. First, my focus is not just on how people use words, but on larger practices of causal inference and explanation in scientific and nonscientific contexts, practices that involve substantial non-verbal components. Second, one of my aims is to make distinctions among different sorts of causal and explanatory claims, distinctions that are often overlooked by those who make such claims. This is not just a matter of describing universally accepted uses. A third and more fundamental difference between my project and conceptual analysis, as conceived above, is that my project focuses on (what I take to be) the purposes or goals behind our practices involving causal and explanatory claims; it is concerned with the underlying point of our practices.<sup>1</sup> Relatedly, my project has a significant *revisionary* or *normative* component: it makes *recommendations* about what one ought to mean by various causal and explanatory claims, rather than just attempting to describe how we use those claims. It recognizes that causal and explanatory claims sometimes are confused, unclear, and ambiguous and suggests how these limitations might be addressed.

In my view, the project of describing our practices having to do with cause and explanation and the project of making recommendations about these



practices are interrelated in complicated ways, and both should be pursued together. (In a similar way, the descriptive and the prescriptive mutually inform each other in other areas of inquiry with a normative component, such as moral philosophy.) On the one hand, anything that qualifies as an account of causation (explanation, etc.), whether descriptive or prescriptive, must be significantly constrained by prior usage, practice, and paradigmatic examples; among other things, these delimit the topic we are talking about. If I tell you that, according to my account, “*C* causes *E*” is coextensive with “*C* is temporally prior to *E*,” you may reasonably respond that whatever I’m talking about, it is not causation. Similarly, if, when a moving billiard ball strikes another and sends it off in a new direction, I deny that this is a causal transaction. On the other hand, although fit with (and illumination of) generally accepted judgments and practice, both in ordinary life and in science, is an important constraint, it does not follow, for the reasons described above, that our only goal should be the description of how ordinary folk (or experts) use words like “cause” and “explanation.”

Where does the normative component of my project come from? It has several sources. First, the purposes behind our practices of making causal and explanatory claims provide a partial metric for assessment of various philosophical proposals about cause and explanation; some proposals fit with those purposes and others do not. Second, there are other sorts of constraints, for example, the epistemological constraints discussed in section 1.7 and constraints having to do with clarity and connection with other concepts and practices. More generally, we introduce concepts (including concepts of cause and explanation) and characterize them in certain ways at least in part because we want to *do* things with them: make certain distinctions, describe certain situations (which usually requires being able to tell whether the concept applies, on the basis of evidence that we have some possibility of getting), calculate with them, use them in proofs or arguments, and so on. Concepts can be well or badly designed for such purposes and we can evaluate them accordingly. Again, previously established uses and practice will be relevant to such an enterprise, because, among other things, they bear on the purposes we want our concepts to serve. But this is not to say that the enterprise reduces just to recording these uses.

The suggestion that it is legitimate or worthwhile to engage in this sort of partially revisionary project will no doubt seem misguided to at least some readers. In fact, however, I contend that such projects are valuable and common parts of most intellectual disciplines, including science and mathematics. Consider the mathematicians who contributed to the rigorization of analysis. Their proposals replaced intuitive, geometrical ideas (about, e.g., “continuity”) with much more precise algebraic definitions. Their investigations also led to the introduction of distinctions whose importance was not previously appreciated (e.g., the distinction between “continuity” and “uniform continuity”). However one conceives of this work, it was definitely not merely conceptual analysis of what the community of mathematicians meant by “continuity” all along. A similar observation holds for much of the work

carried out by nonphilosophers on causation and related matters. For example, the “definitions” of “causal effect” that one finds in writers like Pearl (2000a) and Holland (1986) are not just attempts to describe commonly accepted usage—either of scientists or of ordinary people—and should not be evaluated just on the basis of whether they provide such a description. Again, they have, to be sure, an important continuity with ordinary usage and with scientific practice—otherwise, they could hardly claim to be characterizations of “causal effect” in any sense—but they are also intended as clarifications or regimentations of that usage and practice, introduced with certain purposes in mind (e.g., statistical applications). In fact, I believe that a similar point holds for much of the philosophical literature on cause and explanation: this is also not “conceptual analysis” in the purely descriptive sense described above. We lack an accurate, common, accepted vocabulary for describing the activity carried on in this literature, but it is legitimate and important nonetheless.<sup>2</sup>

### 1.3 The Manipulability Conception of Causal Explanation

I turn now to some brief remarks that are intended to illustrate and motivate the manipulability conception; details come in subsequent chapters. I emphasize that my aim at this point is simply to sketch the general picture I advocate in a very rough (some may think reckless) way; qualifications and refinements are added later.

The manipulability conception plays an important role in the way that scientists themselves think about causal explanation but has received rather less attention from philosophers. The basic idea is nicely illustrated by a contrast drawn between descriptive and explanatory science in a paper by Robert Weinberg (1985) on recent developments in molecular biology. Weinberg tells us that “biology has traditionally been a descriptive science,” but that because of recent advances, particularly in instrumentation and experimental technique, it is now appropriate to think of molecular biology as providing “explanations” and identifying “causal mechanisms.” What does this contrast between description and explanation consist in? Weinberg explicitly links the ability of molecular biology to provide explanations with the fact that it provides information of a sort that can be used for purposes of manipulation and control. New experimental and instrumental techniques have played such a decisive role in the development of molecular biology into an explanatory science precisely because such techniques make it possible to intervene in and manipulate biological systems and to observe the results in ways that were not previously possible. Molecular biologists correctly think that “the invisible submicroscopic agents they study can explain, at one essential level, the complexity of life” because by manipulating those agents it is now “possible to change critical elements of the biological blue print at will” (p. 48).

This passage suggests the underlying idea of my account of causal explanation: we are in a position to explain when we have information that is

relevant to manipulating, controlling, or changing nature, in an “in principle” sense of manipulation characterized in chapter 3. We have at least the beginnings of an explanation when we have identified factors or conditions such that manipulations or changes in those factors or conditions will produce changes in the outcome being explained. Descriptive knowledge, by contrast, is knowledge that, although it may provide a basis for prediction, classification, or more or less unified representation or systemization, does not provide information potentially relevant to manipulation. It is in this that the fundamental contrast between causal explanation and description consists. On this way of looking at matters, our interest in causal relationships and explanation initially grows out of a highly practical interest human beings have in manipulation and control; it is then extended to contexts in which manipulation is no longer a practical possibility. This interest is importantly different from a number of the other interests philosophers have associated with explanation, for example, from our interest in prediction or even in nomically grounded prediction, or from our interest in constructing theories that unify, systematize, and organize in various ways, or that trace spatiotemporally continuous processes. As we shall see, one can have information that is relevant to prediction (including prediction based on generalizations that many philosophers are prepared to regard as laws), or information about spatiotemporally continuous processes, or information that allows for the sort of unification and systemization that many philosophers have thought relevant to explanation, and yet lack the kind of information that is relevant to manipulation on which my account focuses. When this is the case, my view is that one doesn’t have a (causal) explanation. Conversely, one can have information that is relevant to manipulation and hence to explanation, even though one lacks the other features described above. What one needs for manipulation is information about *invariant* relationships, and one can identify invariant relationships even in cases in which one doesn’t know laws, cannot trace spatiotemporally continuous processes, or unify and systematize.

I said above that explanatory information is information that is potentially relevant to manipulation and control. It is uncontroversial, however, that causal relationships exist and that explanation is possible in circumstances in which actual manipulation is impossible, whether for practical or other sorts of reasons. For example, we construct causal explanations of past events and of large-scale cosmological events, and in neither case is manipulation of these phenomena possible. The notion of information that is relevant to manipulation thus needs to be understood modally or counterfactually: the information that is relevant to causally explaining an outcome involves the identification of factors and relationships such that *if* (perhaps contrary to fact) manipulation of these factors *were* possible, this would be a way of manipulating or altering the phenomenon in question. For example, it is currently believed that the explanation (1.2.1) of the mass extinctions at the end of the Cretaceous period has to do with the impact of a large asteroid and the killing effects of the dust it created. Clearly, we cannot now do anything to affect whether this impact occurred, and quite possibly humans could have

done nothing to alter the impact even if they had existed with current levels of technology at the time of impact. My suggestion is that if this explanation is correct, it nonetheless will be true that if it had been possible to alter or prevent the impact, this would have altered the character of or prevented the extinction. Put differently, my idea is that one ought to be able to associate with any successful explanation a hypothetical or counterfactual experiment that shows us that and how manipulation of the factors mentioned in the explanation (the *explanans*, as philosophers call it) would be a way of manipulating or altering the phenomenon explained (the *explanandum*). Put in still another way, an explanation ought to be such that it can be used to answer what I call a *what-if-things-had-been-different question*: the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways. We can also think of this as information about a pattern of counterfactual dependence between explanans and explanandum, provided the counterfactuals in question are understood appropriately. As we shall see, even when actual manipulation is impossible, it is heuristically useful to think of causal and explanatory claims in this way: it both clarifies their content and enables us to understand why they have many of their distinctive features.

On this view, our interest in causal explanation represents a sort of generalization or extension of our interest in manipulation and control from cases in which manipulation is possible to cases in which it is not, but in which we nonetheless retain a concern with what would or might happen to the outcome being explained if various possible changes were to occur in the factors cited in the explanans. If we had been unable to manipulate nature—if we had been, in Michael Dummett's (1964) example, intelligent trees capable only of passive observation—then it is a reasonable conjecture that we would never have developed the notions of causation and explanation and the practices associated with them that we presently possess. Once developed, these notions and practices were then extended to contexts in which actual manipulation was infeasible or impossible. This extension was very natural and perhaps inevitable because, as we shall see in chapter 3, it is built into the notion of a relationship that is usable for purposes of manipulation and control that whether such a relationship holds does not depend on whether the manipulation in question can be actually carried out.

Although it will not be news to historians that the aim of manipulating or controlling nature has played a central role in the development of modern science, this aim has received relatively little attention from philosophers. Most philosophers have distinguished sharply between pure science and applied science or technology, and have regarded explanation as a characteristic aim of pure science and manipulation and control as aims of applied science. To the extent that philosophers have concerned themselves with applied science, they often have seen it as primarily focused on prediction and have failed to appreciate how different prediction is from control. To readers in the grip of this conventional picture of science, my association of our interest

in explanation with our practical interest in control over nature will seem misguided and counterintuitive.

However, a variety of more recent developments in history and philosophy of science and in science studies challenges this sharp distinction between pure science and its application. Part of my intention in writing this book is to contribute to a conception of the role of causal explanation in science that fits with these new developments. I include among these developments recent work in the history of science that emphasizes how concerns with technological application have heavily influenced the content of the more theoretical parts of science (e.g., Smith and Wise 1989; Barkan 1999) and recent work by sociologists, philosophers, and historians on experimentation, which has emphasized in various ways how important our ability to intervene and manipulate nature is in the development of a scientific understanding of nature. Broadly similar ideas can be found in some of the recent philosophical literature on explanation, for example, in Paul Humphreys's (1989) work, with its criticisms of "passive empiricism." On the conception of science that I favor, two aims that are often regarded as quite separate—the "pure science" aim of representing nature in a way that is truthful and accurate and the "applied science" aim of representing nature in a way that permits manipulation and control—are deeply intertwined.

My association of our interest in explanation with our interest in manipulation and control will also seem less surprising when one reflects that a very central part of the commonsense notion of cause is precisely that causes are potential handles or devices for bringing about effects. We find this idea in the manipulability theories of causation defended by writers like Collingwood (1940), Gasking (1955), and von Wright (1971). As we shall see in chapter 2, it is also widely endorsed by social scientists and statisticians, who have shown that this idea can play an important heuristic role in both elucidating the meaning of causal claims and clarifying how statistical evidence can be used to test them. Unfortunately, however, standard philosophical statements of the manipulability theory lead to accounts of causation that are unacceptably anthropocentric and subjectivist. I show in chapter 2 how a manipulability account of causation/explanation can be developed in a way that satisfies reasonable expectations about the objectivity of causal relationships.

#### 1.4 Causal Explanation, Invariance, and Intervention Illustrated

My discussion so far has been rather abstract. It will be useful to have a concrete example in front of us to illustrate some of the ideas to which I have been referring. Consider (1.4.1) a block sliding down an inclined plane with acceleration  $a$  (fig. 1.4.1). What accounts for or explains the motion of the block? The standard textbook analysis proceeds as follows. The block is subject to three forces: a gravitational force due to the weight of the block; a normal force  $N$ , which is perpendicular to the plane; and a force due to

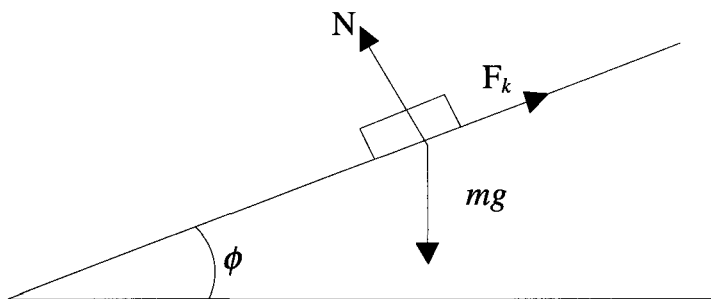


Figure 1.4.1

friction, which opposes the motion of the block. The frictional force  $F_k$  obeys the relationship

$$(1.4.2) \quad F_k = \mu_k N$$

where  $\mu_k$  is the coefficient of kinetic friction. The gravitational force  $F_g$  is directed toward the center of mass of the earth and obeys the relationship

$$(1.4.3) \quad F_g = mg$$

where  $g$  is a constant that represents the acceleration due to gravity near the surface of the earth and  $m$  is the mass of the block.

From the diagram in figure 1.4.1, we see that there is a force directed down the plane of magnitude  $mg \sin \phi$ . The normal force  $N = mg \cos \phi$  and so the frictional force  $F_g = \mu_k mg \cos \phi$ . The net force on the block along the plane is the resultant of these two forces, which is

$$(1.4.4) \quad F_{net} = mg \sin \phi - \mu_k mg \cos \phi$$

The acceleration of the block is thus given by

$$(1.4.5) \quad a = g \sin \phi - \mu_k g \cos \phi$$

How does this explanation work? What is it about the information just conveyed that makes it explanatory? The account I defend takes this and other explanations to provide understanding by exhibiting a pattern of counterfactual dependence between explanans and explanandum—a pattern of counterfactual dependence of the special sort associated with relationships that are potentially exploitable for purposes of manipulation and control. In particular, the above explanation allows us to see how changing the values of various variables in (1.4.2)–(1.4.5) would result in changes in what it is that we are trying to explain: the acceleration of the block. For example, we see from (1.4.5) how changing the angle of elevation of the plane will change the

acceleration of the block: as the angle is increased, the acceleration will be greater. Similarly, the above explanation shows us how, if we were to move the apparatus to a stronger or weaker gravitational field (i.e., if we were to change the value of the “constant”  $g$ ), the acceleration of the block would change. We can also see, by way of contrast, that changing the value of  $m$  will make no difference to the motion of the block. The suggestion that I develop below is that this what-if-things-had-been-different information is intimately connected to our judgments of causal and explanatory relevance: we see whether and how some factor or event is causally or explanatorily relevant to another when we see whether (and if so, how) changes in the former are associated with changes in the latter.

It is a familiar point, however, that not every case of counterfactual dependence corresponds to a causal or explanatory relationship. There is a straightforward interpretation of the notion of counterfactual dependence according to which the joint effects of a common cause such as the reading of a barometer  $B$  and the occurrence/nonoccurrence  $S$  of a storm produced by atmospheric pressure  $A$  are counterfactually dependent on each other; that is, according to which, counterfactuals such as

(1.4.6) If the barometer reading were to fall, a storm would occur

are true. Nonetheless, the barometer reading does not cause or explain the occurrence/nonoccurrence of the storm. Some way must be found of distinguishing the kind of counterfactual dependence that is associated with causal and explanatory relevance from the counterfactual dependence of  $S$  on  $B$ . My solution to this problem appeals to two closely interrelated notions: *intervention* and *invariance*.

For reasons that will become clearer in chapter 2, a manipulability theory is most naturally formulated in terms of variables—quantities or magnitudes that can take more than one value. Causal relationships, of course, have to do with patterns of dependence that hold in the world, rather than with relationships between numbers or other abstracta, but in the interest of avoiding cumbersome circumlocutions, I will often speak of causal relationships as obtaining between variables or their values, trusting that it is obvious enough how to sort out what is meant. Suppose, then, that  $X$  and  $Y$  are variables. The notion of an intervention attempts to capture, in nonanthropomorphic language that makes no reference to notions like human agency, the conditions that would need to be met in an ideal experimental manipulation of the value of  $X$  performed for the purpose of determining whether  $X$  causes  $Y$ . A more precise characterization is provided in chapter 3, but the intuitive idea is that an intervention on  $X$  with respect to  $Y$  changes the value of  $X$  in such a way that if any change occurs in  $Y$ , it occurs only as a result of the change in the value of  $X$  and not from some other source. This requires, among other things, that the intervention not be correlated with other causes of  $Y$  except for those causes of  $Y$  (if any) that are causally between the intervention and  $X$  or between  $X$  and  $Y$ , and that the intervention not affect  $Y$  via a route that fails to go through  $X$ .

Suppose that, in the barometer/storm example, we make use of some process that sets the value of  $B$  at a high or low reading, in a way that is causally and probabilistically independent of the value of  $A$ . This would constitute an intervention on  $B$  with respect to  $S$ . What we expect is that the value of  $S$  would not change if such an intervention were to be performed on  $B$ ; that is, that counterfactuals of the form

(1.4.7) If the value of  $B$  were to be changed as a result of an intervention, then the value of  $S$  would change

are false. My view is that the sorts of counterfactuals that matter for purposes of causation and explanation are just such counterfactuals that describe how the value of one variable would change under interventions that change the value of another. Thus, as a rough approximation, a necessary and sufficient condition for  $X$  to cause  $Y$  or to figure in a causal explanation of  $Y$  is that the value of  $Y$  would change under some intervention on  $X$  in some background circumstances (which may include interventions on other variables besides  $X$  and  $Y$ ; see chapter 2). It is because there is no intervention on  $B$  that will change the value of  $S$  that  $B$  does not cause or figure in an explanation of  $S$ . The difference between interventionist counterfactuals such as (1.4.7) and noninterventionist counterfactuals such as (1.4.6) corresponds, roughly, to the difference between what Lewis ([1973] 1986b) calls nonbacktracking and backtracking counterfactuals. But, as will become clear in chapter 3, this correspondence is only rough; Lewis's view differs from my own both in the judgments it reaches about the truth values of particular counterfactual claims and, more fundamentally, in its basic motivation.

Although interventions on  $B$  (the barometer reading) will not change  $S$  (whether or not the storm occurs), there are at least some interventions or experimental manipulations of the mass of the block, the angle of elevation of the plane, the value of  $g$ , and so on, such that, if they were to occur, the acceleration of the block would change in just the way described by the generalizations (1.4.2)–(1.4.5). In other words, in contrast to the correlation between  $B$  and  $S$ , these equations do correctly tell us how the values of the acceleration would change under (some) interventions. In my view, it is this that accounts for the difference in explanatory status between (1.4.1) (the explanation of the motion of the block) and the “explanation” of the occurrence of the storm in terms of the barometer reading.

The notion of *invariance* is closely related to the notion of an intervention. A generalization  $G$  (relating, say, changes in the value of  $X$  to changes in the value of  $Y$ ) is invariant if  $G$  would continue to hold under some intervention that changes the value of  $X$  in such a way that, according to  $G$ , the value of  $Y$  would change—“continue to hold” in the sense that  $G$  correctly describes how the value of  $Y$  would change under this intervention. A necessary and sufficient condition for a generalization to describe a causal relationship is that it be invariant under some appropriate set of interventions. The generalization describing the correlation between  $B$  and  $S$  in the above example will break down under all interventions on  $B$ : it is a noninvariant generalization. This



corresponds to the fact that  $B$  does not cause  $S$ , and we cannot appeal to this correlation to explain the occurrence or nonoccurrence of the storm.

By contrast, the generalizations (1.4.2) and (1.4.3) in the inclined plane example are invariant under at least a certain range of interventions that change the value of  $N$ . For some such changes,  $\mu_k$  will be (at least approximately) a constant that is independent of  $N$ , and hence (1.4.2) will correctly describe how  $F_k$  will change under this intervention. Similarly, the relationship  $F = mg \sin \phi$  will continue to correctly describe how the component of the weight of the block directed along the plane will change under at least some interventions that change the value of  $\phi$  and of  $m$ .

When a relationship is invariant under at least some interventions in this way, it is potentially usable for purposes of manipulation and control—potentially usable in the sense that, though it may not as a matter of fact be possible to carry out an intervention on  $X$ , it is nonetheless true that *if* an intervention on  $X$  were to occur, this would be a way of manipulating or controlling the value of  $Y$ . Thus, in the above examples, we may control the value of  $F_k$  by controlling the value of  $N$ ; we may manipulate the component of the gravitational force on the block directed along the plane by manipulating  $m$  and  $\phi$ ; and so on. By contrast, one cannot manipulate whether or not a storm occurs by altering the position of a barometer dial. It is in this sense that the account I favor associates causal and explanatory relationships with relationships that tell us about the distinctive patterns of counterfactual dependence that are associated with manipulation and control.

The notion of invariance under interventions is intended to do the work (the work of distinguishing between causal and merely accidental generalizations) that is done by the notion of a *law of nature* in other philosophical accounts. In the above example, although the generalizations (1.4.2) and (1.4.3) represent or describe causal relationships, they are highly problematic candidates for laws of nature. The generalization (1.4.3)  $F = mg$  lacks many of the features standardly ascribed to laws; it holds only approximately, even near the surface of the earth, and fails to hold even approximately at sufficiently large distances from the earth's surface. It is obviously contingent on the earth's having a particular mass and radius. The relationship (1.4.2)  $F_k = \mu_k N$  is, if anything, a less appealing candidate for a law. The textbook from which I took this example goes out of its way not to describe (1.4.2) as a law, but instead describes it as a nonfundamental and only approximate "empirical relationship" (Frautshi, Olenick, Apostol, and Goodstein 1986, pp. 179ff). The particular value of the coefficient of kinetic friction between a pair of surfaces is the net result of a large number of extremely complicated contact forces and depends, in ways that are still not well understood from the perspective of fundamental theory, on the detailed characteristics of the two surfaces and will change if the characteristics of those surfaces are altered. Thus, even if (1.4.2) correctly describes the frictional force on the block for some particular experimental setup, we could easily disrupt that generalization by, for example, greasing the surface of the plane or abrading it with sandpaper. This is very different from the behavior of paradigmatic laws of nature.

As (1.4.2) and (1.4.3) illustrate, whether or not a generalization is invariant is surprisingly independent of whether it satisfies many of the traditional criteria for lawfulness, such as exceptionlessness, breadth of scope, and degree of theoretical integration. This is in large measure a consequence of the way in which invariance is defined: for a generalization to be invariant, all that is required is that it be stable under *some* (range of) changes and interventions.<sup>3</sup> It is not required that it be invariant under *all* possible changes and interventions. Thus, the generalization (1.4.2), describing the relationship between frictional and normal force, is invariant as long as it would continue to hold under some interventions that change the value of  $N$ . Its invariance is not undermined by the fact that there are other interventions on  $N$  (e.g., increasing  $N$  to a very large value) or other sorts of changes (greasing the contact surface) under which (1.4.2) would break down. I argue below that the notion of invariance is better suited than the notion of lawfulness for capturing the distinctive features of many of the generalizations that describe causal relationships and figure in explanations. The notion of law imports many features that generalizations like (1.4.2) do not seem to possess. Calling (1.4.2) a law immediately places us in the dialectically awkward position of needing to explain why that characterization is appropriate, even though (1.4.2) lacks many of the features usually ascribed to laws, whereas describing (1.4.2) as invariant generates no such difficulties. Rather than thinking of all causal generalizations as laws, I suggest that we should think of laws as just one kind of invariant generalization.

## 1.5 Some Varieties of Causal Claims

I have been claiming that causal relationships share a common feature: they are invariant relationships that are potentially exploitable for purposes of manipulation and control. But within this broad genus there are additional distinctions that can be drawn among varieties of causal relationships and among dimensions along which causal relationships may vary. For example, one way in which the explanation (1.4.2) differs from the explanation (1.3.1) of the extinction of the dinosaurs is that the former is not (at least as I have presented it) an explanation of any specific episode involving a block sliding down a plane. Instead, what is explained is a generic pattern in the motion of blocks down planes—what James Bogen and I (Bogen and Woodward 1988) call a *phenomenon*. By contrast, what is explained in (1.3.1) is a specific episode of extinction. Causal claims or explanations of particular events like (1.3.1) are often called *token causal claims* or *singular causal explanations* and, as we shall see, they have a number of distinctive features. They share with other varieties of causal explanation the general feature that they are to be understood in terms of counterfactuals about what will happen under interventions, but the relevant counterfactuals differ in various ways from those associated with explanations like (1.4.2). Similarly, type-level causal claims like those in (1.4.2) may either have to do with what I call in chapter 2 *total*

or *net* causes or with *direct* or *contributing* causes, and these also will differ in the counterfactuals with which they are associated. In general, we may illuminate the content of different sorts of causal claims and causal notions by describing the different hypothetical experiments with which they are associated.

There is yet another important dimension along which causal claims may differ. Consider the claim

(1.5.1) Depressing the gas pedal on my car causes it to accelerate.

Assuming that my car is functioning normally, the manipulationist view that I advocate judges that this is a true causal claim and one to which I might appeal to explain why my car has accelerated on some particular occasion. Nonetheless, as the econometrician Haavelmo (1944) observed decades ago, there is an obvious sense in which it is explanatorily shallow in comparison with the sort of theory of the internal mechanism of the car that might be provided by an automotive engineer; I don't have any very deep understanding of why my car moves as it does, if I know only (1.5.1). This example illustrates the point that causal explanations differ in the degree or depth of understanding they provide. One would like a theory of causal explanation to provide some insight into what makes one causal explanation deeper or better than another. The theory I develop attempts to do this by tracing differences in explanatory depth to differences in the degree of invariance of the generalizations to which the explanation appeals (i.e., differences in the range of changes and interventions over which those generalizations are stable) and to differences in the range of what-if-things-had-been-different questions that the generalization answers.

## 1.6 Causal Explanation as a Practical Activity

I said above that explanations and causal inference pervade our lives. I mean by this that these are widespread, everyday activities in which most human beings, including people in other cultures quite different from our own, engage. They are, if you like, "natural" human activities, rather than activities that only scientists or philosophers with a taste for extravagant metaphysics engage in. Few cultures have developed the systematic procedures for investigating nature that we think of as science, but all cultures, including those in the distant past, have been curious about causal and explanatory relationships and have accumulated a great deal of causal knowledge of a mundane sort; for example, that exposure to fire or heat causes pain and tissue damage, that the impact of a moving rock can cause another object to break or move, and that crops require water to grow. I take this fact to suggest (though I readily acknowledge that it hardly conclusively establishes the correctness of) several additional ideas about how to understand our explanatory practices. First, it suggests that the construction of explanations and the acquisition of causal knowledge must have, at least sometimes, some practical point or payoff. There must be some benefit, other than the satisfaction of idle curiosity, that is

sometimes provided by these activities. One way of interpreting different theories of causation and explanation is to think of them as providing different answers to the question of what this benefit might be. As already intimated, the theory that I defend takes the distinctive benefit associated with causal and explanatory knowledge to have to do with manipulation and control.

Second, we should expect there to be some sort of continuity between the everyday practices and kinds of causal/explanatory knowledge that are present in all human societies and the more systematic and sophisticated practices and kinds of causal/explanatory knowledge that characterize contemporary science. We should expect continuity on both substantive and methodological levels. On a substantive level, and in contrast to the tendency among some philosophers to regard all folk or commonsense beliefs as fundamentally mistaken, I see causal explanation in science as building on and requiring causal knowledge of a more mundane, everyday sort. If we did not have prior causal knowledge of the sort possessed by craftspeople and artisans about how to manipulate the natural world, or if we did not know various simple causal truths (e.g., massive objects fall when unsupported, solid macroscopic objects cannot pass through one another, gases and liquids diffuse unless surrounded by solid containers), we could not construct instruments or carry out or interpret even simple experiments. Similarly, on a methodological level, we should expect causal explanations in different areas of science to share at least some structural features with causal explanations in more ordinary contexts. We should see scientists who construct such explanations as attempting to satisfy some of the same explanatory goals and interests as people who construct explanations in ordinary life, but as achieving those goals by appealing to knowledge that is more rigorous, detailed, and systematic. Explanation in science does not give us something that is fundamentally different in kind from explanation in more ordinary contexts, but rather, as it were, a better version of the latter.

Given that we should expect some continuity between causal explanation in ordinary life and the sorts of explanation provided by sophisticated scientific theories, what form should we expect this continuity to take? There are at least two possibilities. One is to use structures and categories characteristic of some forms of scientific explanation to understand explanation in ordinary life. This is the approach taken by most philosophers, including defenders of the *DN* model and adherents of unificationist models of explanation such as Philip Kitcher (1989; see chapter 8). For example, as the account of the motion of the block down the plane (1.4.1) illustrates, explanations in science often (although by no means always) rely on explicit chains of deductive reasoning in a way that explanations in nonscientific contexts often do not. Thus, one might try to capture the continuity between ordinary and scientific explanations by understanding the former as having the structure of an “implicit” deductive argument or as explanatory in virtue of “tacitly” invoking or relying on such arguments—arguments that are explicit or overt in the case of scientific explanations. Similarly, some explanations in physics appeal to generalizations that are (or so I argue below) appropriately described as laws of nature. One might try to capture the continuity between such

explanations and other sorts of explanation by arguing that all causal explanations “tacitly” rely on “backing” laws even when they do not explicitly appeal to generalizations at all or do not appeal to any generalization that might naturally be described as a law.

For a variety of reasons described in more detail below, I find this approach unpersuasive. The “backing” relationship that allegedly holds between garden-variety causal claims and laws is difficult to make clear. Moreover, there are crucial features of the content of causal claims and causal explanations that are not captured by the “instantiation of laws” picture just described. I argue below that the *DN* view of the relationship between ordinary and scientific explanation gets matters exactly backwards. All human cultures have produced causal explanations, but the notion of a deductively valid argument and the notion of a law of nature are complex and sophisticated products of a very specific intellectual and scientific tradition. Rather than trying to understand all varieties of causal explanation in terms of these specialized categories, we should instead begin with a more general notion of causal explanation, understood in manipulationist terms, and then attempt to understand explanations that appeal to explicit chains of deductive reasoning and laws of nature as one specific variety within this genus.

### 1.7 Accounts of Causation and Explanation Can Be Illuminating without Being Reductionist

There is another set of issues that deserves comment. There is a very widespread tendency in philosophical discussions of causation and explanation to assume that any interesting account of these notions must be “reductive.” Just what this means is rarely made clear, but I take the general idea to be that concepts like “cause” and “explanation” belong to an interrelated family or circle of concepts that also includes notions like “law,” “physical possibility,” and other modally committed notions. (The circle may also include “counterfactual dependence,” depending on how this notion is understood.) An account is reductive if it analyzes concepts in this family solely in terms of concepts that lie outside of it. It is also usually assumed that the acceptable concepts in terms of which a reductive analysis might be framed must satisfy certain epistemological and metaphysical constraints of an “empiricist” stripe; for example, it is assumed that the analysis should appeal only to notions that are “actualist” or nonmodal in the sense that they have to do with what actually happens or will happen, rather than what must or can happen. So-called regularity theories of law and causation are among the most familiar theories that are reductive in this sense, but there are other examples as well, such as Lewis’s counterfactual theory. It is frequently assumed that any account of causation or explanation that fails to be reductive will be “circular” and hence unilluminating.

The account that I present is not reductive, and I am skeptical that any reductive account will turn out to be adequate. However, little if anything in

my positive account turns on whether this skepticism is correct. Indeed, I would be delighted if someone were able to show how the nonreductive characterizations of cause and explanation that I provide might be replaced by reductive characterizations. By contrast, it is crucial to my argument that an account of causation and explanation can be worthwhile and illuminating without being reductive. Of course, the only really convincing way of showing this is to actually produce the account in question and to allow the reader to see that it is illuminating. Nonetheless, some general remarks at this point may be helpful in allaying misgivings about the nonreductive character of the theory that follows.

It is perfectly true that some nonreductionist theories of causation/explanation (e.g., *c* causes/explains *e* if *c* produces or generates *e*), with no further account of “production” or “generation,” are completely unilluminating. But not all nonreductive theories are trivial in the way just illustrated. One way in which nonreductive theories can be interesting and controversial rather than trivial and empty is by virtue of conflicting with other reductive or nonreductive theories and suggesting different assessments of particular explanations. For example, according to the manipulationist account of explanation that I defend, explanations that involve action at a distance or otherwise fail to trace spatiotemporally continuous causal processes nonetheless can be genuinely explanatory. Theories such as Salmon’s causal/mechanical model reach the opposite conclusion (see chapters 3 and 8). Which of these approaches is more plausible seems quite independent of whether either achieves a successful reduction.

A second point that is worth keeping in mind is this: even if we opt for a nonreductive account according to which some notions in the circle of concepts that includes “cause,” “explanation,” and so on are explained in terms of other notions in that circle, we still face many nontrivial choices about exactly how the various notions in this circle should be connected with or used to elucidate one another—choices that can be made in more or less defensible ways. For example, although I offer a nonreductive treatment of the kinds of counterfactuals that are relevant to elucidating “cause,” “law,” and “explanation,” I also argue that the counterfactuals on which the philosophical tradition has tended to focus in elucidating these notions are the wrong counterfactuals for this purpose. Again, the correctness of this claim seems completely independent of questions of reduction. As another illustration of the same point, I argue below that to elucidate certain kinds of causal claims, including claims about direct causal relationships and singular causal claims, one must appeal to counterfactuals with complex antecedents—counterfactuals that describe what will happen under combinations of manipulations or interventions, rather than under single manipulations. These sorts of counterfactuals have rarely been used by philosophers to elucidate causal claims, although one may think of them as implicit in some treatments of causation devised by nonphilosophers. Whether these are the right counterfactuals to look at in elucidating these causal claims or whether other counterfactuals would be more appropriate is again completely independent of the issue of reduction.

My general view is that in their enthusiasm for reductive accounts, philosophers have often misdescribed or oversimplified the content of the causal and explanatory claims they have hoped to reduce. We need more careful description of just what such claims say (and of the regularities and counterfactuals associated with them). Only after this has been done should we investigate what sorts of reductions are possible.

Yet another point is that even in the absence of a fully reductive account of causation and explanation, it may be possible to test or elucidate the content of particular causal and explanatory claims and show how they can be tested by appealing to other particular causal/explanatory claims and noncausal information such as correlational claims. In other words, we may be able to test or elucidate the claim that  $C_1$  causes  $E_1$  by appealing to our antecedent knowledge that some *other* causal claim (e.g.,  $C_2$  causes  $E_2$ ) is true, along with other noncausal information, perhaps about correlations. The theory I propose has this sort of structure: it holds that we may test or elucidate the claim that  $C$  causes  $E$  by appealing to what will happen to  $E$  under an intervention on  $C$ . The notion of an intervention is itself a causal notion—among other things, it involves the idea of an intervention variable  $I$  that *causes* a change in  $C$ —but the causal relationships to which we need to appeal in characterizing what it is for  $I$  to be an intervention on  $C$  are different from the causal relationship between  $C$  and  $E$  that we are trying to elucidate. An account of this sort is not reductive, because it doesn't explain causation in terms of the concepts that lie outside of the circle of concepts to which "cause" belongs, but it is not viciously circular in the way that explaining "cause" in terms of a primitive notion of "production" would be.

## 1.8 Epistemic Constraints on Explanation

A theory of causal explanation should also satisfy plausible epistemological constraints. It is true and important that we need to respect the distinction between issues about the content of causal or explanatory claims—issues about what such claims mean or what they say—and epistemological issues about how we test such claims or determine whether they are true. However, in our enthusiasm for this distinction, we should not overlook the equally important need to tell an integrated and plausible story about how these two sorts of issues fit together. In particular, our theory of the content of causal and explanatory claims should be accompanied by some epistemological story that makes it understandable how human beings can sometimes learn whether claims with that content are true or false from evidence that is actually available to them. This story should enable us to understand how (or to what extent) widely accepted procedures for testing causal and explanatory claims—for example, controlled experimentation and the causal modeling techniques discussed in chapter 7—work. To put the point negatively, if our theory of what causal and explanatory claims say leaves it a complete mystery how we ever find out whether claims with that content are true, or if procedures like

controlled experiments that we ordinarily think of as testing such claims have no discernible bearing on whether claims with that content are true, this is an indication that something is fundamentally wrong. Either our theory about what such claims say or our views about how we should determine whether such claims are true or both need to be rethought.

Causal relationships are features of the world: they are “out there” in nature. By contrast, explanation is an activity carried out by humans and conceivably by some other animals, having to do with the discovery and provision of *information*, information about causal relationships. This leads, I argue, to an additional epistemic constraint on explanation that has no counterpart if our concern is just with causal claims. The constraint is, roughly, that explanatory information must be epistemically accessible information. It must be information that can be recognized, surveyed, or appreciated—in short, information that can contribute to understanding. The significance of this constraint is explored in chapter 4.

## 1.9 Desiderata

Drawing together the various strands of this discussion, I offer by way of summary the following nonexhaustive list of goals/constraints for a theory of causation and explanation, ordered from less to more controversial.<sup>4</sup>

The theory should be descriptively adequate in the sense that it captures relevant features of paradigmatic explanations in science and ordinary life. It should give us some insight into how such explanations work in the sense that it identifies the features or structures in virtue of which they convey understanding.

If the theory recognizes different varieties or sorts of causal explanation (as the theory that I propose does), it should show us what these have in common: why it is that they all count as species of the genus “causal explanation.”

The theory should allow us to evaluate explanations. It should help us to distinguish between better and less good explanations, and it should enable us to understand the grounds on which such normative assessments are made. It should distinguish causal and explanatory claims from claims that are merely descriptive.

The theory should elucidate the successes and failures of previous philosophical theories of causal explanation; it should solve problems that are not adequately dealt with in previous theories.

The theory should have adequate epistemological underpinnings. If the theory tells us that an explanation works by conveying certain information or by possessing a certain structure, then there should be some plausible accompanying epistemological story that makes it clear how people who use the explanation can learn about this information or structure, how they can check whether the claims embedded in the explanation are correct, and so on. More generally, if an explanation provides understanding by conveying certain information, then this information should be epistemically accessible to those



who use the explanation (cf. chapter 4). Relatedly, the theory should enable us to make sense of widely accepted procedures for testing causal and explanatory claims.

The remainder of this book is organized as follows. Chapters 2 and 3 explore various notions of causation within a broadly manipulationist framework. Chapters 4 and 5 focus on the notion of causal explanation, using ideas about causation developed in previous chapters. Chapter 6 discusses the notion of invariance and its relationship to standard philosophical ideas about lawfulness. Chapter 7 applies the ideas about causation and explanation defended in previous chapters to the so-called structural equations or causal modeling literature. Chapter 8, the concluding chapter, compares my account of causation and explanation to two well-known alternative accounts, the *causal mechanical* model of Wesley Salmon (1984) and the *unificationist* model of Philip Kitcher (1989).

## Causation and Manipulation

In the previous chapter, I suggested that causal and explanatory claims are informed by our interest as practical agents in changing the world.<sup>1</sup> In particular, I claimed that it is heuristically useful to think of explanatory and causal relationships as relationships that are potentially exploitable for purposes of manipulation and control. This idea is stressed in traditional manipulability theories of causation such as those developed by Gasking (1955), Collingwood (1940), and von Wright (1971). It also underlies recent theories that focus on the connection between causation and agency, such as Menzies and Price (1993). To a large extent, however, philosophical discussion has been unsympathetic to manipulability theories: it is claimed both that they are unilluminatingly circular and that they lead to an implausibly anthropocentric and subjectivist conception of causation. This negative assessment among philosophers contrasts sharply with the widespread view among statisticians, theorists of experimental design, and many social and natural scientists that an appreciation of the connection between causation and manipulation can play an important role in clarifying the meaning of causal claims and understanding their distinctive features.

Illustrations of this second view can be found in almost any textbook on experimental design. For example, in their highly influential book on quasi-experimentation, Cook and Campbell (1979) write: “*The paradigmatic assertion in causal relationships is that manipulation of a cause will result in the manipulation of an effect. . . . Causation implies that by varying one factor I can make another vary*” (p. 36; emphasis in original). A similar point of view is expressed by the statistician Paul Holland (1986) in the form of a slogan (“No causation without manipulation”) and by David Freedman (1997) in the following passage: “Causal inference is different [from description], because a change in the system is contemplated: for example, there will be an intervention. Descriptive statistics tell you about the correlations that happen to hold in the data: causal models claim to tell you what will happen to *Y* if you change *X*” (p. 116). Very similar ideas can be found among economists. For example, Kevin Hoover (1988) writes in his survey of neoclassical economics that the following “definition of cause is widely acknowledged”:

A causes B if control of A renders B controllable. A causal relation, then, is one that is invariant to interventions in A in the sense that if someone

or something can alter the value of  $A$  the change in  $B$  follows in a predictable fashion. (p. 173)

In a similar vein, Gary Orcutt (1952) writes that “we see that the statement that . . .  $Z_1$  is a cause of  $Z_2$  is just a convenient way of saying that if you pick an action that controls  $Z_1$ , you will also have an action that controls  $Z_2$ ” (p. 307).

As I noted in chapter 1, similar claims are common among molecular biologists. We can also find physical scientists who express related views. For example, in the course of an argument that superluminal signals are incompatible with special relativity, Roger Newton (1993) appeals first to the idea that the notion of a signal is a causal notion—a signal from  $A$  to  $B$  involves transmission of a causal influence of some sort for  $A$  to  $B$ —and then to what is essentially a manipulability account of what it is to transmit causal influence:

What constitutes a signal? It is clear from our arguments above that a signal from  $A$  to  $B$  is anything that can activate a switch at  $B$  on command from  $A$ . This would necessarily include any transmission of energy or of information. However, there are some motions that are not of that nature. For example, if you hold your finger in front of a light bulb and observe the shadow cast on a distant screen, the speed of the shadow, as you move your finger, is proportional to the distance to the screen. If that distance is large enough, in principle the shadow can move as fast as you please. However, as it rushes from a point  $A$  to a point  $B$ , both on the screen, it cannot be used to activate a switch at  $B$  on command from  $A$ . Therefore, a superluminal speed of such a shadow would not violate the relativity-causality based prohibition. There are many other speeds of phenomena in physics that are also free from this limitation.

The joker in the argument of the last paragraph is the phrase “on command from.” It may stimulate you to contemplate the role that our belief in “free will” plays in what we mean by causality. We certainly believe that we are free to send any message we want, and to observe its effect. Indeed, all scientific experimentation is based on the assumption that we are free to wiggle something here and to record the response over there. (pp. 139–40)

I take Newton’s idea to be roughly this: if there is a causal process leading from  $A$  to  $B$ , then it should be possible for the right sort of “command” at  $A$  to “activate a switch” at  $B$ , which is just to say that a suitable manipulation at  $A$  should produce a change in  $B$ . Thus, if the state of the shadow at  $A$  could cause changes in the state of the shadow at  $B$ , this would mean that it would be possible in principle to intervene (of your own “free will”) to alter the state of the shadow at  $A$  and in this way to change its state at  $B$ . Because any such change in the shadow at  $A$  would not, as a matter of empirical fact, result in a change at  $B$ , Newton holds that it is wrong to think of the movement of the shadow from  $A$  to  $B$  as involving a direct causal relationship or the transmission of causal influence.

Philosophical readers may be tempted to dismiss remarks of the sort just quoted as reflecting a naïve lack of awareness of the complexities surrounding the notion of causation and the problems facing a manipulationist approach. This would be uncharitable. Even if we put aside remarks like Newton's about causation in physics, understanding causation and causal inference is a central, if controversial, concern of disciplines like statistics, experimental design, and econometrics, and researchers in these disciplines encounter many of the same problems and complications that draw the attention of philosophers. Broadly manipulationist accounts of causation are reflected in substantial portions of the *practice* of these disciplines and not just in the comments of researchers about that practice. To the extent that this practice is successful, manipulationist approaches deserve a detailed exploration, rather than peremptory dismissal. As we shall see, many objections that have been thought to be fatal can be addressed in a natural way within a manipulationist framework.

The very different assessments (on the part of philosophers and nonphilosophers) of the manipulability conception of causation generate a puzzle. Are the writers quoted above simply mistaken in thinking that focusing on the connection between causation and manipulation can tell us something valuable about causation? Does the widespread invocation of something like a manipulability conception among practicing scientists show that the usual philosophical criticisms of manipulability theories of causation are misguided? One of my themes in what follows is that the different assessments of manipulability accounts of causation in and outside of philosophy derive from the different goals or aspirations that underlie the versions of the theory developed by these two groups. Philosophical defenders of the manipulability conception have typically attempted to turn the connection between causation and manipulability into a reductive analysis: their strategy has been to take as primitive the notion of manipulation (or some related notion like agency or bringing about an outcome as a result of a free action), to argue that this notion is not itself causal (or at least does not presuppose all of the features of causality we are trying to analyze), and to then attempt to use this notion to construct a noncircular reductive definition of what it is for a relationship to be causal. Philosophical critics have (quite reasonably) assessed such approaches in terms of this aspiration (i.e., they have tended to think that manipulability accounts are of interest only insofar as they lead to a noncircular analysis of causal claims) and have found the claim of a successful reduction unconvincing. By contrast, statisticians and other nonphilosophers who have explored the link between causation and manipulation generally have not had reductionist aspirations. Instead, their interest has been in unpacking what causal claims mean and how they figure in inference by tracing how they link up with the notion of manipulation, but without suggesting that the notion of manipulation is itself a causally innocent notion. As such, they have followed the general nonreductionist strategy recommended in chapter 1: that of trying to elucidate the concept of causation by tracing its interconnections with or locating it in a "circle" of interrelated concepts, but without claiming to analyze the concepts in this circle in terms of concepts that live entirely outside it.

I agree with the philosophical critics that the reductionist version of the manipulability theory is unsuccessful. Indeed, attempts to analyze causation in terms of manipulation turn out to be “circular” not just in the obvious sense that for an action or event *I* to constitute a manipulation of a variable *X*, there must be a causal relationship between *I* and *X*, but in other, more subtle and interesting ways as well: for *I* to qualify as a manipulation in the sense relevant to understanding causation, *I* must not just be causally related to *X* but must be an event or process with a very special kind of causal structure, and to characterize this structure we must make extensive use of causal notions. (Following the lead of several other writers, e.g., Meek and Glymour 1994; Pearl 2000a, I call a manipulation having the right sort of structure an *intervention*.) In fact, it is largely because the notion of manipulation/intervention has this sort of fine structure—a structure that is unexplored in traditional manipulability theories—that working out the connection between causation and manipulation turns out to be interesting and nontrivial rather than banal and obvious. I also agree with the critics that reductively inclined manipulability theories lead to an unacceptably subjectivist and anthropocentric treatment of causation. In fact, as we will see, the impulse toward reduction virtually forces one in such directions. However, by giving up the goal of reduction, we may develop a manipulability conception that is genuinely illuminating but not unacceptably anthropomorphic or subjectivist. I believe that it is something very like this conception that the writers quoted above, who think that the connection between causation and manipulation has something interesting to teach us about causation, have in mind.

## 2.1 Motivation

As a preliminary motivation, let me begin with a question that is not often asked in philosophical treatments of causation: What is the point of our having a notion of causation (as opposed to, say, a notion of correlation) at all? What role or function does this concept play in our lives? An important part of the appeal of a manipulability account of causation is that it provides a more straightforward and plausible answer to this question than its competitors.

A natural inclination of many philosophers, when confronted with this question, is to stress the role of disinterested intellectual curiosity and to downplay or dismiss the role of more practical considerations. It will be said that we care about distinguishing between causal and merely correlational relationships because knowledge of causal relationships plays a central role in explanation and understanding, especially in “pure” or “theoretical” science. Although I do not deny that there is such a thing as disinterested intellectual curiosity, this response is unsatisfying. For one thing, it leaves it mysterious why our theoretical interests and curiosity express themselves in the particular way they do. For example, it does not explain why our theoretical interests are not satisfied simply by knowledge of correlations.

More important, a very substantial body of research has made it clear that a disposition to distinguish between (some) causal and noncausal sequences is widely shared among humans and many nonhuman animals, emerges early in development, and in some cases is remarkably fast and efficient, requiring relatively few trials or only limited evidence. In a classic series of experiments, Garcia, Ervin, and Koelling (1966) gave various types of food to rats and then exposed them to radiation several hours later that made them nauseated. The rats learned to avoid the foods that were followed by illness on the basis of a single trial. By contrast, when Garcia et al. exposed the rats to avoidable light and sound signals followed by radiation-induced nausea, the rats did not learn to avoid these signals even on repeated trials, although they did learn very quickly to avoid such signals when they were paired with electric shocks. To put the matter in a way that, although tendentious, is also suggestive: the rats behaved as though they assumed that if they became nauseated this must have been caused by something they ate (and hence that nausea could be avoided by altering what they ate), but that exposure to noises and lights were not the sorts of thing that could cause nausea. Needless to say, these causal hypotheses are correct in the rat's natural environment, although not in the artificial environment of Garcia et al.'s laboratory.

Similar predispositions to distinguish between causal and noncausal sequences are found among human infants. In the "launching" experiments carried out by John Leslie and his colleagues (e.g., Leslie and Keeble 1987), researchers observed the reactions of young infants (twenty-seven weeks) to simulated normal collisions in which one moving object strikes another, with the result that the second appears to move. They then compared these reactions with infant reactions to causally abnormal sequences in which the first object stops some distance from the second and then the second spontaneously appears to move away. Infants habituate more quickly to the former display than the latter, a result that is taken to show that they are more surprised by the second sort of behavior or that it violates their expectations in a way that the first does not. Young infants react similarly to other sorts of causally impossible behavior: solid objects that appear to pass through each other on collision, objects that remain suspended in midair when their supports are removed, and so on (Baillergon, Kotovsky, and Needham 1995). Such "expectations" are conjectured to play an important role in learning the skills and causal knowledge associated with physical manipulation of objects in the infant's environment. (cf. Plotkin 1997). Experiments on older children (e.g., four years) show a much richer set of causally charged assumptions that play an important role in categorization, classification, and generalization. Contrary to what is sometimes suggested, there is little evidence that causal expectations or assumptions about the behavior of familiar physical objects vary greatly across different cultures or that non-Western cultures have a concept of causality that is fundamentally different from our own (cf. Sperber, Premack, and Premack 1995).

Although we need not interpret these results as showing that causal knowledge is "innate,"<sup>2</sup> either in humans or other animals, it is plausible to

conclude that the development of these abilities has some practical benefit, connected to such goals as survival and reproduction. As Plotkin (1997) puts it, “We humans come into the world primed or predisposed to learn certain features of the causal structure of that world” (p. 189), and this must be because learning about the causal structure of the world has some practical payoff for us. “Disinterested intellectual curiosity” is not a good explanation of why rats and infants have the abilities described above. The question I want to raise thus can be framed as follows: What does causal knowledge enable us (or other animals) to achieve with respect to such practical goals as survival and reproduction that other kinds of knowledge does not? In asking this question, I do not mean to suggest that causal knowledge serves only these goals and no others. My point is, rather, that any acceptable account of causation should explain why causal knowledge is sometimes practically useful and what its practical utility consists in.

Broadly speaking, philosophical theories of causation fall into two main camps with respect to this question. Some theories, call them “impractical,” provide accounts that fail to make it understandable how knowledge of causal relationships has any practical utility at all. Other theories, call them “practical,” do provide (or are naturally interpretable as providing) accounts that connect casual knowledge with some goal that has practical utility, with different theories providing different accounts of what that goal consists in. My suggestion is that, because of the considerations described above, we have strong *prima-facie* reasons to prefer practical to impractical theories, or at least to regard impractical theories as incomplete and in need of supplementation. I argue below that among practical theories, the manipulability account is the most promising.

One example of an impractical theory, or so I argue in chapter 3, is David Lewis’s ([1973] 1986b) counterfactual theory of causation. As another example, consider the conserved quantity theory of Wesley Salmon (1984, 1994) and Philip Dowe (2000).<sup>3</sup> According to this theory, causal interactions involve the intersection of two or more causal processes and the exchange of a conserved quantity, such as energy or momentum, as when two billiard balls collide. By contrast, noncausal interactions do not involve the exchange of energy and momentum. There are reasons to doubt that this is an extensionally adequate theory, in the sense that it correctly distinguishes between causal and noncausal interactions (see chapter 8). However, even putting this issue aside, the conserved quantity theory, as it stands, fails to explain the practical point of (the benefit that would follow from) the ability to distinguish between interactions that involve the exchange of energy and momentum and those that do not. Needless to say, animals, small children, and many adults do not possess the concepts of energy, momentum, and conservation laws and are not motivated by a disinterested intellectual concern to classify interactions into those that do and those that do not involve the transfer of energy and momentum. If the ability to distinguish between causal and noncausal interactions, as these are understood in the conserved quantity theory, is practically useful, this must be because this distinction tracks or

coincides with (perhaps only roughly) some other distinction, which is more directly beneficial. For example, if we were to assume that the distinction between those interactions that involve the transfer of energy and momentum and those that do not coincides with the distinction between those types of interactions that can be used for successful prediction or manipulation and those that cannot, there would be no mystery about why the distinction is practically useful. However, as we shall see (chapter 8), this assumption is not correct. Moreover, if it were correct, it would immediately raise the question of why we should not take notions having to do with prediction and control, rather than notions having to do with energy conservation, as the most promising route to understanding causation.

There are many different accounts of causation that qualify as practical. For reasons of space, I focus on just two of these: “regularity” theories and the manipulationist account I favor. A simple and naïve deterministic version of a regularity theory says that  $C$ s cause  $E$ s if and only if there are additional background circumstances  $K$  such that all  $C$ s in  $K$  are followed by  $E$ s. A simple probabilistic version of a regularity theory says that  $C$ s cause  $E$ s if and only if  $C$  raises the probability of  $E$  in some (or, alternatively all) suitable background conditions  $B_i$ , that is, iff  $P(E/C \cdot B_i) > P(E/-C \cdot B_i)$  for some or all  $B_i$  (cf. Eells 1991). One of the attractions of such theories is that they provide a straightforward account of the practical value of causal knowledge: such knowledge is valuable because it can improve our ability to predict. On both of the above versions, if we know that  $C$  has occurred, then knowing that (i)  $C$ s cause  $E$ s will improve our ability to predict whether  $E$  will occur relative to a situation in which we do not know (i). If  $E$  is an event that can affect an organism’s chances of survival and reproduction, there is no puzzle about why such predictive information may be useful.

The obvious problem with prediction as the (or a) distinctive goal of (or rationale for) causal inquiry is that is overbroad: a concern with prediction doesn’t explain why we make the distinctions we do between causal and noncausal relationships. Suppose that  $E_1$  does not cause  $E_2$  and  $E_2$  does not cause  $E_1$  but that  $E_1$  and  $E_2$  are correlated because they are both effects of a common cause  $C$ . Knowing that  $E_1$  and  $E_2$  are correlated, we can use the occurrence of one to predict the occurrence of the other, just as we can use the occurrence of  $C$  to predict the occurrence of both  $E_1$  and  $E_2$ , but we nonetheless distinguish among these predictive relationships, labeling some as causal and others as noncausal. Why do we bother to make these distinctions if the point of causal knowledge is merely to predict? Similarly, from the occurrence of an effect (e.g., smoke, red spots on the body) we can often infer the occurrence of its cause (fire, measles), but we think of the latter as causing the former and not vice versa. Indeed, inferences from effect to cause are often more reliable than inferences from cause to effect: from the footprints on the beach, I can reliably infer that someone walked on it, but from knowledge that someone has walked on the beach, I can reliably infer little about the existence of footprints. Examples of this sort are often advanced as counterexamples to regularity theories, but we can also take them to strongly suggest that the practical payoff



associated with causal knowledge is not just an improved general ability to predict. For that purpose, knowledge of correlations alone would suffice.

There are ways of complicating regularity approaches to deal with the examples just described,<sup>4</sup> but rather than pursuing these, I want to explore the answer to the question of practical utility that is associated with a manipulability account. This is that the distinctive practical payoff of causal knowledge has to do with its usefulness for manipulation and control. One can think of this idea as a subspecies of (or refinement of) the generalized prediction rationale described above. 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 on the basis of observations of situations in which she or the other agent have not (yet) acted. Consider a well-known illustration. Nancy Cartwright ([1979] 1983a, 1983b) describes a letter she received from TIAA-CREF, which provides life insurance to college teachers. The letter contains the following passage:

It simply wouldn't be true to say, "Nancy L. D. Cartwright . . . if you own a TIAA life insurance policy you'll live longer."

But it is a fact, nonetheless, that persons insured by TIAA do enjoy longer lifetimes on average, than persons insured by commercial insurance companies that serve the general public. (1983b, p. 22)

Purchase of TIAA life insurance (*I*) is correlated with increased longevity (*L*) in the U.S. population. However, as Cartwright notes, this fact does not by itself show that purchasing such insurance is a means or effective strategy for increasing the probability that one will live longer; that it is a way of manipulating longevity. In fact, it seems likely that the correlation between *I* and *L* is not due to *I*'s being a cause of *L* or to *L*'s being a cause of *I*, but is rather due entirely to some set of common causes of both *I* and *L*. That is, the correlation arises because decisions to purchase insurance are influenced by characteristics that also tend to cause increased longevity. For example, the academics who purchase TIAA insurance on average have a higher income than people in the general population and may for this reason have better access to health care. They may also be better informed about health-related matters and have less stressful jobs, and this may have an independent effect on longevity. If the correlation between *I* and *L* arises entirely as a result of some set of common causes of *I* and *L*, that correlation will not be one that an agent can exploit in such a way as to increase life span through the purchase of insurance. If we consider a group of people who are exactly matched for whatever other factors influence longevity, some of whom decide to purchase TIAA insurance and others of whom do not, or alternatively, a random sample of people who have not purchased TIAA insurance, some of whom are then randomly assigned insurance and others of whom are not, we would not expect there to be a correlation between *I* and *L* in these populations. In this sense, the correlation between *I* and *L* will not be stable or invariant under efforts to use *I* to control or manipulate *L*.

This example is representative of a characteristic problem faced by human beings and many other animals. One may learn, through passive observation, that two variables *A* and *B* are correlated. However, this fact by itself tells one nothing about whether one can, by acting so as to change or manipulate *A*, also change *B*, and such information often will be of crucial practical import. I take the guiding idea of a manipulability approach to causation to be that lying behind the distinction we make between causal relationships and mere correlations is a concern to distinguish between, on the one hand, a relationship between *A* and *B* that can be used to manipulate (in the sense that if it were possible to manipulate *A*, this would be a way of changing *B*) and, on the other hand, a correlation that will simply disappear when we attempt to manipulate *B* by manipulating *A*. Thus, the question of whether purchase of insurance causes or is merely correlated with increased longevity is practically important because only in the former case can an agent manipulate longevity by purchasing insurance.<sup>5</sup>

As another illustration, consider an early hominid who observes that rocks tend to move when struck by other rocks of roughly equal size in cases in which such impacts occur “naturally,” without human intervention. This information will be predictively useful regardless of whether such impacts cause or are merely correlated with the movement of the second rock. Nonetheless, the question of whether the impact causes or is merely correlated with the movement is far from idle: it matters practically because it has implications for whether the hominid can manipulate whether the second rock moves by throwing or pushing the first rock into it. If the correlation between impact and movement is like the correlation between purchase of insurance and longevity, then trying to move the second rock by pushing the first into it will be futile. By contrast, if the impact causes the movement, then, if it is possible for the hominid to make the first rock strike the second in the right way, this will be a way of making the second rock move—potentially a very useful piece of information. Similarly, the question of whether a rat’s consumption of a food item causes or is merely correlated with subsequent nausea is of crucial practical importance because it bears on the question of whether the rat can avoid nausea by refraining from eating the item. If food consumption and nausea are merely correlated, there is no point in trying to avoid nausea by not eating the food.

In both of these cases, the possibility that the relationships in question are merely correlational seems so unlikely that it is easy to overlook the point that a separate inferential step is required to move from the claim that two kinds of events are correlated to a claim about how one will respond under manipulation of the other. In other cases, this second inferential step will be controversial and the need to justify it obvious. Numerous studies (e.g., Coleman and Hoffer 1987) show that children attending private schools have higher academic achievement, as measured by scores on standardized tests, than children attending public schools, even when one controls for factors like family income. This correlation may be due in part to the fact that attending a private school causes increased scholastic achievement, but it is also possible

that it results entirely from the operation of various difficult-to-measure attitudinal variables (the importance parents attach to education, etc.) that serve as common causes of both the decision to attend a private school and of high scholastic achievement. There is considerable disagreement among experts about which of these alternative hypotheses is correct, and a parent or policymaker who observes the correlation between private schooling and academic achievement cannot simply assume that academic achievement can be boosted by private schooling. A manipulationist approach to causation seems to nicely capture the differences in the practical import of these two alternative causal hypotheses; they have very different implications for whether it would make sense for parents to attempt to improve their child's achievement via private schooling.

As several writers note (Plotkin 1997; Gopnik, Glymour, Sobel, Schulz, Kushnir, and Danks forthcoming) the distinction between information about correlations and information about relationships that will support manipulations is reflected in the distinction that is commonly drawn between two kinds of conditioning: classical or Pavlovian conditioning and instrumental or Skinnerian conditioning. In classical conditioning, an organism learns an association between two sorts of events that are outside its control. In Pavlov's original experiments, dogs form an association between the ringing of a bell and the provision of food, as shown by the fact that they begin to salivate just when the bell is rung. As Plotkin remarks, classical conditioning puts the experimental subjects in the position of the investigator envisioned in classical empiricism: able to observe correlations in the world but not able to intervene in it. By contrast, in instrumental conditioning, what is learned is an association between some behavior produced by the experimental subject and a consequence of the behavior. For example, rats may learn an association between pressing a certain lever and the provision of food; after such learning, they behave as though they understand that pressing the lever causes a food pellet to be provided. It is information learned in instrumental conditioning that is directly relevant to manipulation and control. According to a manipulationist conception of causality, it is this second sort of information that is information about causal relationships. As Dickenson and Shanks (1995, p. 23) put it, it is "the capacity to control rather than just react to the environment that provided the impetus for the evolution of a mind and a nervous system capable of representing causality" (quoted in Plotkin 1997, p. 199).

A second set of considerations that seem to support a broadly manipulationist approach to causation is rooted in the role of experimentation in causal inference. It is very widely believed by scientists and statisticians that when it is possible to carry them out, appropriately designed experiments often can provide better evidence for causal claims than passive observation. For example, there is general agreement that if we wish to test the hypothesis that ingestion of a new drug causes recovery from some disease, it is better to employ an appropriately designed randomized experiment than to simply make the drug available to all to wish to take it and observe how the incidence of recovery among those who take the drug compares with those who do not.

A manipulationist approach to causation explains the role of experimentation in causal inference in a straightforward way: experimentation is relevant to establishing causal claims because those claims consist in, or have immediate implications concerning, claims about what would happen to effects under appropriate manipulations of their putative causes. In other words, the connection between causation and experimentation is built into the very content of causal claims. By contrast, this connection is much less clear on alternative accounts of causation. For example, on the Salmon–Dowe theory, the difference between those drugs that cause recovery and those that do not must have something to do with the distinction between those interactions that involve some appropriate transfer of quantities like energy and momentum and those that do not. Defenders of that theory thus owe us an account of why the results of controlled experiments with drugs turn out to be a very reliable way of distinguishing between these two sorts of interactions, whereas other sorts of studies are not, even though in their overt design the experiments have no obvious connection with energy and momentum transfer, and even though it is far from clear what it means to say (and still less clear that it is true) that the distinction between those drugs that cause recovery and those that do not coincides with the distinction between those drugs that transfer energy and momentum to recovery and those that do not. I do not claim that it is impossible to provide such an account, but merely that it needs to be provided, and that it is not obvious how to do this while drawing only on ideas underlying the Salmon–Dowe account.

Of course, causal inference based on real-life experiments is by no means always unproblematic or error-free. There are many different ways that real-life experiments can mislead about causal relationships, some of which are discussed below. Nor, of course, do I mean that one can learn about causal relationships *only* through experiments, or that experimentation is *always* superior to passive observation as a way of finding out about causal relationships. There are cases in which one can learn more about causal relationships (assuming one is willing to make certain other assumptions) from observation or from a combination of observation and experiment than from practically possible experiments alone, and there are many cases in which, for moral or practical reasons, one must rely on nonexperimental evidence to reach causal conclusions.<sup>6</sup> A plausible manipulability theory will not deny that reliable causal inference on the basis of nonexperimental evidence is possible, but rather, suggests a specific way of thinking about such inferences: we should think of them as an attempt to determine (on the basis of other kinds of evidence) what the results of a suitably designed hypothetical experiment or manipulation would be without actually carrying out this experiment. For example, for moral and political reasons, among others, one cannot carry out experiments in which some children are randomly assigned to public and others to private school and their subsequent academic careers observed. Nonetheless, it is illuminating to think of attempts to infer from nonexperimental data the effects of private schooling on achievement as attempts to predict what the results of such an experiment would be without

actually doing it. As we shall see, an important part of the heuristic value of a manipulability account is that by clearly spelling out what it is that we are trying to establish when we make causal claims (i.e., what the result of a certain hypothetical experiment would be), it forces us to think in a disciplined and specific way about the nonexperimental evidence and other assumptions that would be required to justify such claims. A closely related point is that thinking about causal claims in terms of hypothetical experiments enables us to distinguish among different possible interpretations of those claims: different possible interpretations will be associated with different possible hypothetical experiments. (See section 3.2 for additional discussion.)

There are other respects in which a manipulability theory is philosophically appealing. As readers of the philosophical literature will know, appeals to "intuitions" about whether various relationships are causal or explanatory have played an important role in discussion of these subjects. Such appeals tend to be unsatisfactory in several respects. To begin with, people's intuitions often disagree. More fundamentally, without some account of the point or purpose that is served by causal claims or the larger rationale that underlies our willingness to regard certain relationships as causal and others as non-causal, it is hard to see how to resolve such disagreements or even to see what is at stake in deciding to resolve them in one way rather than another. A manipulability theory addresses this problem. If, as a manipulability theory suggests, the point of distinguishing causal claims from claims about correlations that do not reflect direct causal connections is (roughly) to distinguish those relationships that are potentially exploitable for purposes of manipulation from those that are not, then we have a principled reason for thinking that causal relationships must possess certain features and not others. For example, we have a principled reason for denying that it is built into the notion of causation that if  $X$  causes  $Y$ , then  $X$  must be connected to  $Y$  via a spatiotemporally continuous process (see chapter 3). We also have a principled reason for distinguishing among a variety of different causal concepts (direct causes, contributing causes, total causes) that often have not been clearly distinguished.

Moreover, by drawing our attention to the underlying rationale behind various particular causal judgments, manipulability theory also helps to make the whole notion of causation seem less puzzling and metaphysically mysterious. Philosophers who have thought that causal notions are metaphysically excessive (or import problematic metaphysical baggage) have been influenced by many different arguments, but one important consideration has been the suspicion that we could accomplish all of our legitimate purposes if we had no causal knowledge or beliefs at all, but instead made do with weaker forms of knowledge (such as knowledge of correlations). Part of the attraction of a manipulability theory is that it helps us to see that there are legitimate and important purposes, purposes that are rooted in our interests as practical agents and that we could hardly fail to have if we are to go on living. The distinction between causal and merely correlational relationships is not just the product of an impulse to engage in woolly, otherworldly metaphysics, as some philosophers (e.g., van Fraassen 1980) suggest.

Even if one is skeptical that such a manipulability account will adequately capture all of the features that are assigned to the concept of causation in ordinary life and the various sciences, the quotations with which I began this chapter should make it clear that at least some writers and research traditions simply *mean* by “causal relationship” something like “relationship potentially exploitable for purposes of manipulation.” Moreover, regardless of whether this is “all there is” to causation, the contrast between such relationships and those that are merely correlational is real and important in both science and ordinary life. It is surely worthwhile to try to understand what this contrast consists in and its role in scientific and ordinary practice. Those who doubt that a complete account of causation will emerge from the connection between causation and manipulation may think of what follows as an investigation of the notion of a relationship that is exploitable for purposes of manipulation—a notion that is interesting and important in its own right (even if it does not fully coincide with “our” notion of causation) and that deserves more philosophical attention than it has hitherto received.

Finally, we may note that the considerations just described also suggest certain awkward questions for approaches that fail to take seriously the connection between causation and manipulation. Consider, for example, the common suggestion that although there may be something right about the idea that in practical, experimental, or “applied” science contexts causal relationships are relationships that are potentially exploitable for purposes of control, it would be wrong to think of causal relationships in “pure” or “theoretical” science in this way (cf. Hausman 1998, pp. 164ff). Does this mean that there are two quite distinct notions of causation, one appropriate for practical contexts and the other for theoretical contexts? What happens when, because of technological advances, it becomes possible to carry out an experimental manipulation to test a causal claim in what was previously a nonexperimental area of science or to apply what was previously purely theoretical causal knowledge to manipulate nature in new ways? Do causal claims that were previously part of theoretical science undergo a fundamental change in meaning in such cases? Surely, a far more plausible view is that the meaning or role of causal claims is the same in both contexts. Even in purely theoretical contexts, causal claims should be understood as telling us about the results of hypothetical manipulations; it is just that we cannot, at least at present, carry out these manipulations. If, at a later time, it becomes possible to test such claims experimentally, we don’t fundamentally alter their meaning. Rather, all that changes is that we now have an opportunity to assess the correctness of the predictions they make about what would happen under various manipulations by actually carrying out the manipulations in question. However, these predictions about what would happen under manipulations were part of the content of the causal claims, and were correct or incorrect, all along. This, at any event, is the view that I defend in this chapter.

I fully concede that the considerations just described do not show that it is possible to construct an illuminating version of a manipulability theory. Their role instead is intended to be motivational. They ought to convince us that the

connection between causation and manipulation is worth exploring more systematically, and that it is a mistake to suppose that the idea that there is such a connection is a naïve bit of anthropomorphism that plays no role in scientific practice. On the contrary, there are substantial traditions in many areas of science that think of causal relationships as relationships that can be used to manipulate.

## 2.2 Graphs and Equations as Devices for Representing Causal Relationships

I turn now to the task of formulating a defensible version of a manipulability theory. First, however, some stage setting is in order, beginning with a few words about the scope of my project in this and the following chapter and what is distinctive about it. My aim is to give an account of the content or meaning of various locutions, such as  $X$  causes  $Y$ ,  $X$  is a direct cause of  $Y$ , and so on, in terms of the response of  $Y$  to a hypothetical idealized experimental manipulation or intervention on  $X$ . This project is heavily influenced by ideas in Judea Pearl's *Causality* (2000a) and also owes a great debt to ideas in Spirtes, Glymour, and Scheines ([1993] 2000).

As I understand Pearl's enterprise, it takes as primitive various qualitative notions of causal dependence (e.g., the notion of  $X$  being directly causally relevant to  $Y$ ), defines the notion of an intervention by reference to this notion, and then shows us how to calculate or estimate various quantitative causal notions (such as the magnitude of the total effect of  $X$  on  $Y$ ) in terms of this framework. Spirtes et al. are, by their own account, less interested in issues about what various sorts of causal claims mean and focus instead on problems of causal inference or discovery from statistical data. By contrast, I have nothing to say about issues having to do with calculating quantitative magnitudes, estimation, identifiability, or causal inference. Instead, my enterprise is, roughly, to provide an account of the meaning or content of just those qualitative causal notions that Pearl (and perhaps Spirtes et al.) take as primitive. Because my project is semantic or interpretive, and is not intended as a contribution to practical problems of causal inference, it requires a somewhat different notion of intervention than the notion assumed by Pearl (the notion that I employ is closer to the notion found in Spirtes et al.). The causal locutions that I seek to define are also different in some respects from those on which Pearl focuses. For example, where Pearl is interested in characterizing a quantitative notion of the *direct effect* of  $X$  on  $Y$  (e.g., Pearl 2000a, pp. 126ff), my interest is in using a manipulationist framework to characterize what it is for  $X$  to be a *direct cause* of  $Y$  and more generally to explain what is meant when an equation or directed graph of the sort described below is interpreted causally.

However, I also do not want to exaggerate the originality of what follows. The use of directed graphs to represent causal relationships and the adoption of a broadly interventionist framework for understanding causation (including

the crucial idea that interventions “break arrows” or “wipe out” equations in which the variable intervened on occurs as a dependent variable) can be found in Spirtes et al. and in Pearl, who in turn get the general framework from a tradition in econometrics that goes back to such writers as Haavelmo (1944), Frisch ([1938] 1995), and Strotz and Wold (1960). My characterization of various causal notions, such as the notion of direct cause and the notion of actual cause, is also very heavily indebted to Pearl, as readers will see. Indeed, to a considerable extent, what I have done is simply to place those ideas (or my versions of them) within a more general philosophical framework and to relate them to concerns that may be more familiar to philosophers. Because Pearl’s work (and, to some extent, the work of Spirtes et al.) as well as the tradition in econometrics on which they draw is less well-known to philosophical readers than it deserves to be, and the strengths of an interventionist approach to causation underappreciated by philosophers, there is perhaps some merit in such an enterprise.

As already intimated, on a manipulability account of causation, it is most perspicuous to think of causal relationships as relating *variables* or, to speak more precisely, as describing how changes in the value of one or more variables will change the value of other variables. This is also the way that many scientists think about causal relationships. My primary (but not exclusive) focus is on developing an account of a relatively simple and straightforward kind of causal claim: claims of the form *X* is causally relevant to *Y*, where *X* and *Y* are (or are represented by) variables.<sup>7</sup> To avoid cumbersome circumlocutions and following what has become a well-established usage, I use the expression “*X* causes (or is a cause of) *Y*” interchangeably with “*X* is causally relevant to *Y*,” but I caution that this represents a departure from ordinary usage in some respects (see below).

Intuitively, variables are properties or magnitudes that, as the name implies, are capable of taking more than one value. Values (being red, having a mass of 10 kilograms) stand to variables (color, mass) in the relationship of determinates to determinables. Values of variables are always possessed by or instantiated in particular individuals or units, as when a particular table has a mass of 10 kg. Many of the familiar examples of so-called property causation discussed in the philosophical literature may be understood as relationships between two-valued or binary variables, with the variables in question taking one of two values, depending on whether the properties in question are instantiated or not. Thus, the claim that ingestion of aspirin causes recovery from headache may be understood as asserting a relationship between the values of a variable *A*, representing whether or not aspirin is ingested, and the values of a variable *H*, representing whether or not relief from headache occurs. However, variables need not be two-valued; they may also assume many values or be continuous. For example, the claim that  $F = -kX$ , where *X* is the extension of a spring and *F* the restoring force it exerts, may be understood as asserting that a causal relationship exists between the values of two variables, each of which may take a range of real numbers as its value. Just what causal claims relating variables mean is addressed at some length



below, but the basic idea is simple: the claim that  $X$  causes  $Y$  means that for at least some individuals, there is a possible manipulation of some value of  $X$  that they possess, which, given other appropriate conditions (perhaps including manipulations that fix other variables distinct from  $X$  at certain values), will change the value of  $Y$  or the probability distribution of  $Y$  for those individuals. In this sense, changing the value of  $X$  is a means to or strategy for changing the value of  $Y$ . (As we shall see, this characterization gives us a rather weak or minimal notion of causation between variables, which can be strengthened.) Causal relationships between variables thus carry a hypothetical or counterfactual commitment: they describe what the response of  $Y$  *would be* if a certain sort of change in the value of  $X$  were to occur.

The claim that  $X$  is causally relevant to  $Y$ , where  $X$  and  $Y$  are variables, is often described as a *type-causal* claim. Such claims contrast (in several different respects) with *token-causal* or *actual cause* claims, claims to the effect that some particular token event has caused another (e.g., a specific episode of aspirin ingestion by Smith caused a specific episode of headache recovery). Section 2.7 describes how the account of causal relationships between variables developed below can be extended to token-causal claims.

Although there is a distinction between type- and token-causal *claims*, it does not follow that there are two *kinds* of causation—type and token—or that in addition to token-causal relationships involving particular values of variables possessed by particular individuals, there is a distinct variety of causal connection between properties or variables that is independent of any facts about token-causal relationships. In my view, a claim such as “ $X$  is causally relevant to  $Y$ ” is a claim to the effect that changing the value of  $X$  instantiated in particular, spatiotemporally located individuals will change the value of  $Y$  located in particular individuals. Thus, the truth of a claim such as (S) “Smoking causes lung cancer” depends on relationships that do or would obtain (under appropriate manipulations) at the level of particular individuals, even though it is also true that (S), as it stands, does not tell us anything about the causes of any particular episode of lung cancer (such as Smith’s cancer, diagnosed on such and such a date), or indeed that any individual either smokes or develops lung cancer. In accord both with common usage and the informal characterization of “ $X$  is causally relevant to  $Y$ ” given above, I assume that the claim (S) would be true, even if no one were to smoke, as long as it is the case (as it presumably is) that manipulating whether some particular human being (or beings) smoke will change whether they develop (or the probability of their developing) lung cancer. In other words, it is the response schedule of lung cancer to smoking that matters for the truth of (S), even though the values of the those variables {*smokes, does not smoke*} and {*develops lung cancer, does not develop lung cancer*} are always realized in particular individuals.

There is another feature of the locution “ $X$  is causally relevant to  $Y$ ” that is a potential source of confusion. Both in ordinary and philosophical usage, claims of the form “ $C$  causes  $E$ ” are often understood as relating the occurrence of *types of events* (with other terms, such as  $\neg C$  and  $\neg E$  used to refer to the nonoccurrence of events of these types). In other words,  $C$  is used to refer to

something like a *value* of a variable rather than the variable itself, with  $\neg C$  being used to refer to the other possible value of that variable. Relatedly, “ $C$  causes  $E$ ” is understood to mean something like “ $C$  is a *positive* (as opposed to a negative) causal factor for  $E$ ” or “ $C$ s produce or favor (as opposed to prevent or interfere with)  $E$ s.” The claim that “ $X$  is causally relevant to  $Y$ ” is *not* equivalent to or interchangeable with the claim that  $X$  is a positive causal factor for  $Y$ . In particular, it does not follow from the claim that  $X$  is causally relevant to  $Y$  that each particular value of  $X$  causes or is a cause of (in the sense of being a positive causal factor for) each particular value of  $Y$ . The notion of  $X$  being causally relevant to  $Y$  is a broader, more general notion, corresponding roughly to one factor’s being *either* positively or negatively relevant or of mixed positive and negative relevance to another. Thus, the variable  $X$  that takes the values  $\{\textit{smokes, does not smoke}\}$  should be distinguished from the event or event type of smoking and the claim that the variable  $X$  is causally relevant to the variable  $Y$   $\{\textit{develops lung cancer, does not develop lung cancer}\}$  is true, even though smoking causes lung cancer but not the absence of lung cancer, in the “is a positive causal factor for” sense of cause. Moving from the claim that the variable  $X$  is causally relevant to the variable  $Y$  to the claim that type of event  $C$  causes (in the positive causal factor sense) type of event  $E$  requires additional, more specific information about the character of the relationship between  $X$  and  $Y$ —in particular, about which values of  $X$  are associated with which values of  $Y$ . From the perspective of a manipulability theory, among the advantages of the “variable  $X$  is causally relevant to variable  $Y$ ” locution are its greater generality and clarity. It is often unclear how to apply the “event  $C$  causes event  $E$ ” locution to nonbinary variables and, as we shall see, the attempt to force all causal relationships among variables into a binary framework creates a number of difficulties.

In this chapter, my main focus is on providing an account of causation that applies to deterministic contexts, with a few side remarks about the indeterministic case. On a manipulability theory, if  $X$  is a deterministic cause of  $Y$ , then a set of associated counterfactuals specifying how  $Y$  would change under manipulation of  $X$  (and possibly other variables as well) will be true. If  $X$  is instead an indeterministic or probabilistic cause of  $Y$ , then (depending on the details of the case) there may be no associated true counterfactuals of this form, but there will be associated counterfactuals specifying how the probability distribution of  $Y$  or the probability that  $Y$  will assume some value will change under manipulation of  $X$  (and possibly other variables). It is of course desirable to have an account of causation that applies to indeterministic contexts as well, but indeterministic causation involves a number of subtleties that are best addressed separately. The obvious strategy for generating a manipulability theory of indeterministic causation is to replace the references in the deterministic version to manipulations of the value of  $X$  that change the value of  $Y$  with references to manipulations of  $X$  that change the probability distribution of  $Y$ . Subsequent chapters explore this idea in more detail.

As the previous paragraphs suggest, the account of claims of the form  $X$  causes  $Y$  that I propose is restricted to cases in which the relationship between  $X$  and  $Y$  is general or reproducible in the sense that  $Y$  exhibits some sort of

systematic response when the same changes in the value of  $X$  are repeated, at least in the right circumstances. In the deterministic case,  $Y$  always changes in the same way under the same kind of manipulation of  $X$ , at least for some range of manipulations for individuals within some range of background circumstances. When causation is indeterministic, reproducibility instead may be understood to cover a range of possibilities: it may be that for some range of background circumstances, the same kind of manipulation of  $X$  always changes the probability of  $Y$  by the same value or, more weakly, that repetition of a kind of manipulation of  $X$  in the same background circumstances always produces either an increase (or decrease) in the probability of  $Y$ , although perhaps by different amounts on different occasions. At least this much reproducibility or regularity in the response of  $Y$  to  $X$  is often taken to be built in to the meaning of type-causal claims. Arguably, it is also assumed when we take manipulating  $X$  to be a means or strategy for manipulating  $Y$ , which is the central notion that I will be trying to explicate. Alternatively, one may think of reproducibility as an additional, independent restriction on the scope of our discussion. Without reproducibility, the counterfactual claims on which my account relies will not be obviously true.

In what follows, I make use of two devices to represent causal relationships. A *directed graph* is an ordered pair  $\langle \mathbf{V}, \mathbf{E} \rangle$  where  $\mathbf{V}$  is a set of vertices that serve as the variables representing the relata of the causal relation and  $\mathbf{E}$  a set of directed edges connecting these vertices. A directed edge from vertex or variable  $X$  to vertex or variable  $Y$  means that  $X$  *directly causes*  $Y$ . For now, I will rely on the reader's intuitive understanding of this notion; it is characterized more precisely in section 2.3 below. The basic idea is that  $X$  is a direct cause of  $Y$  if and only if the influence of  $X$  on  $Y$  is not mediated by any other variables in the system of interest  $\mathbf{V}$  in the following sense: there is a possible manipulation of  $X$  that would change the value of  $Y$  (or the probability distribution of  $Y$ ) when all other variables in  $\mathbf{V}$  are held fixed at some set of values in a way that is independent of the change in  $X$ .<sup>8</sup> I assume that if  $X$  is a direct cause of  $Y$ , then  $X$  is a cause of  $Y$ , but that the converse of this claim is false. A sequence of variables  $\{V_1 \dots V_n\}$  is a *directed path* or *route* from  $V_1$  to  $V_n$  if and only if for all  $i(1 \leq i < n)$  there is a directed edge from  $V_i$  to  $V_{i+1}$ .  $Y$  is a *descendant* of  $X$  if and only if there is a directed path from  $X$  to  $Y$ . If  $Y$  is a descendant of  $X$ , then  $X$  is an *ancestor* of  $Y$ . The direct causes of  $X$  are also said to be the *parents* of  $X$ .<sup>9</sup> As we will see below, for  $X$  to be a cause of  $Y$ , it is necessary but not sufficient for  $X$  to be an ancestor of  $Y$ .

It will also be useful to represent causal relationships by means of systems of equations, as in the structural equations literature in econometrics. As explained above, I generally focus on the deterministic case, in which each endogenous variable  $Y$  (i.e., each variable that represents an effect) may be written as a function of all and only those variables that are its direct causes; that is, if variables  $X_1 \dots X_m$  are all of the direct causes of  $Y$ , then  $Y$  may be written as

$$(2.2.1) \quad Y = F_Y(X_1 \dots X_m)$$

There are two features of equations like (2.2.1) that deserve special emphasis. First, they do not just represent patterns of association; instead, each deterministic equation is understood as encoding counterfactual information about how  $Y$  would change under manipulations of its direct causes. In particular, if we write  $X_i = x_i$  to indicate that the variable  $X_i$  takes or is assigned the value  $x_i$  as the result of an idealized manipulation/intervention, then we may think of the equation (2.2.1) as telling us what the value of  $Y$  would be if the variables representing the direct causes of  $Y$  were assigned the values  $X_1 = x_1, X_2 = x_2 \dots X_m = x_m$ , as a result of such manipulations: in such a case, the value of  $Y$  would be  $y = F_Y(x_1 \dots x_m)$ . Thus, if  $Y = 3X_1 + 4X_2$  and  $X_1$  is assigned the value  $X_1 = 2$  and  $X_2$  is assigned the value  $X_2 = 5$  by manipulations, then  $Y$  should assume the value  $Y = 26$ .

Second, unless explicitly indicated otherwise, I always understand equations such as (2.2.1), as well as the graphs associated with them, as applying at the level of (i.e., as purported descriptions of the behavior of) whatever individuals are regarded as possessing particular values of the variables figuring in those equations and graphs. The equations and graphs describe how the value of  $Y$  possessed by those individuals would change in response to the values of  $X_1 \dots X_n$  possessed by those individuals. Thus, (2.2.1) is *not* to be understood as merely a claim about the average or aggregate response of the value of  $Y$  to various average values of  $X_1 \dots X_n$  in some heterogeneous population, the behavior of the members of which may not be characterized at the individual level by (2.2.1).

When underlying causal relationships are deterministic but not all of the direct causes of  $Y$  are known or measured, it may be possible to write  $Y$  as a function of all of its known direct causes,  $X_1 \dots X_k$ , plus a so-called error term  $U$  that enters additively into the equation for  $Y$  and represents the combined influence of all the other unknown direct causes of  $Y$  that are not explicitly represented in the equation:  $Y = F(X_1 \dots X_k) + U$  (see chapter 7). When  $Y, X_1 \dots X_k$ , and  $U$  are random variables with a well-defined joint probability distribution, the presence of the error term makes possible nontrivial conditional probabilities that are strictly between 0 and 1, even when the underlying structure is deterministic. When causal relationships are genuinely indeterministic, the relevant equations specify how the probability distribution of  $Y$  will change under manipulation of the right side variables representing direct causes in each equation.

What is the relationship between the representation of causal relationships by means of systems of equations and their representation by means of directed graphs? When we draw a directed graph with arrows from  $X_1 \dots X_m$  into  $Y$ , we convey the information that  $Y$  is some function of  $X_1 \dots X_m$  and that all of the variables  $X_1 \dots X_m$  are essential in the sense that for each such variable  $X_i$ , there is some combination of values of the others such that some change in  $X_i$  will change  $Y$ . However, the graph does not further specify what this function is. Thus, when we explicitly specify the function or equation relating  $Y$  to its direct causes (e.g.,  $Y = 3X_1 + 4X_2$ ), we convey more information than if we merely draw a graph with arrows from  $X_1$  and  $X_2$  directed

into  $Y$ . Unlike the directed graph, the explicit form of the equation specifies exactly how changing  $X_1$  and  $X_2$  will change  $Y$ . By contrast, the corresponding directed graph says simply that there is some change in  $X_1(X_2)$  that, given some value of  $X_2(X_1)$  will change  $Y$ .

To forestall confusion, it also will be helpful to distinguish between two different sorts of causal structures that involve the operation of direct causes. Consider first a structure in which some dependent variable of interest  $Z$  is an additive function of two other variables  $X$  and  $Y$  which represent the direct causes of  $Z$ :

$$(2.2.2) \quad Z = aX + bY$$

where  $a$  and  $b$  are fixed coefficients. The corresponding directed graph is figure 2.2.1.

In this case, regardless of the value at which we fix one of the variables, interventions on the other make the same fixed contribution to  $Z$ . It follows from (2.2.2) that regardless of the value at which  $X$  is held fixed, an intervention that changes  $Y$  by amount  $\Delta y$  will always change  $Z$  by the same amount  $b\Delta y$ . By way of contrast, suppose that  $O$  is a variable that takes the values 1 or 0 depending on whether oxygen is present or absent,  $S$  is a variable that takes the values 1 or 0 depending on whether a short circuit occurs, and  $F$  is a variable that takes the values 1 or 0 depending on whether a fire occurs. We may then represent the claim that  $O$  and  $S$  are direct causes of  $F$  by means of the equation:

$$(2.2.3) \quad F = S \cdot O$$

or by means of the directed graph in figure 2.2.2.

Although this graph has exactly the same structure as the graph in figure 2.2.1, equation (2.2.3) tells us that a change in  $S$  will not have the same effect on  $F$  regardless of the value at which  $O$  is set. Unlike  $X$  and  $Y$  in (2.2.2), there is an “interaction” between  $S$  and  $O$  with respect to  $F$ . When oxygen is present, changing the value of  $S$  from 0 to 1 will lead to the occurrence of a fire, but when oxygen is absent, a similar change in  $S$  will lead to no change in  $F$ . The reference in the informal characterization of direct causation above to interventions on the cause variable that change the value of the effect variable for *some* (not necessarily all) values of other variables is meant to accommodate

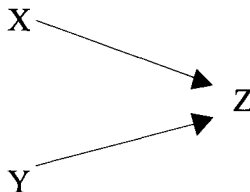


Figure 2.2.1

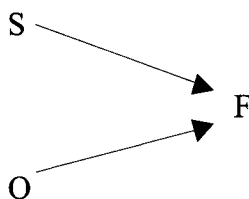


Figure 2.2.2

these observations. For  $S$  to qualify as a direct cause of  $F$ , we require only that there be some value of  $O$  (namely,  $O = 1$ ) such that some change (produced by an intervention) in the value of  $S$  (from 0 to 1) will change the value of  $F$ . Because (2.2.2) and (2.2.3) share the same directed graph, we can see that the graphical representation is insensitive to the difference between causal relationships that involve interactions and those that do not.

### 2.3 Direct, Total, and Contributing Causes

That causal relationships are relationships that can be exploited for manipulation is a vague idea that needs to be made more precise along many different dimensions. One initial question is whether we should take the fact that a relationship can be used for purposes of manipulation to be necessary or sufficient or both for it to be causal. Consider the following proposal for a sufficient condition for causation, which I label **(SC)** for future reference:

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

This is paralleled by the following necessary condition for causation:

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

Before exploring these proposals, a number of clarifying comments are in order. First, what is it to “change” the value of a variable? If we think of an intervention as the taking of a particular value by an intervention variable  $I$ , then  $I$  changes the value of  $X$  and, in doing so, changes the value of  $Y$  if and only if there are values,  $x_1, x_2$ , of  $X$  with  $x_1 \neq x_2$  such that  $I$ 's assuming some value  $i_1$  causes  $X$  to assume the value  $x_1$  and  $I$ 's assuming some distinct value  $i_2$  causes  $X$  to assume the value  $x_2$  and there are values  $y_1, y_2$  of  $Y$  associated with  $x_1$  and  $x_2$  such that  $y_1 \neq y_2$ . In other words, if  $Y = F(X)$ , describes the functional relationship between  $X$  and  $Y$ , there are distinct values of  $X$ ,  $x_1$  and  $x_2$ , and values of  $Y$ ,  $y_1$ , and  $y_2$  such that  $y_1 = F(x_1) \neq y_2 = F(x_2)$ .

Second, why the restriction in clause (ii) of **SC** (and **NC**) to a single intervention on  $X$ ? Suppose that  $X$  does not cause  $Y$ , but that whenever an intervention occurs that changes  $X$  to some value, a second intervention also occurs that changes  $Y$  to some particular value. Then  $Y$  changes systematically under interventions on  $X$ , even though there is no causal relationship between  $X$  and  $Y$  in violation of **SC**.<sup>10</sup> One way of excluding this sort of possibility is to require that  $Y$  change under a single intervention on  $X$  and no others.

Third, a central difficulty with both **SC** and **NC** is to say what “possible” in clause (i) means. For example, if “possible” is taken to mean something like “within the present technological powers of human beings,” then **NC** has the obviously unacceptable consequence that  $X$  cannot cause  $Y$  when human beings lack the power to manipulate  $X$ . On this reading, it would follow that, for example, causal claims about the past and about the influence of the moon on the tides automatically come out false. I address issues concerning the sense in which interventions must be “possible” in chapter 3. For now, I simply assume that the relevant notion of possibility must be broad enough to allow causal claims to come out true even when human beings lack the relevant powers to manipulate.<sup>11</sup>

Finally, we should note that if (**SC**) is to be even *prima facie* plausible, we need to impose some restrictions on what sorts of changes in  $X$  count as “interventions.” To return to the example in chapter 1, suppose that, in a certain region, changes in atmospheric pressure ( $A$ ) are a common cause of the occurrence of storms ( $S$ ) and of the reading ( $B$ ) of a particular barometer, that  $B$  does not cause  $S$  and  $S$  does not cause  $B$ , and that  $A$  is the only common cause of  $B$  and  $S$ . It is clear that there are ways of changing  $B$  that will be associated with a corresponding change in  $S$  even though  $B$  does not cause  $S$ . For example, if we change  $B$  by changing  $A$ , or by means of some causal process that is perfectly correlated with changes in  $A$ , then  $S$  will also change, but this would not establish that  $B$  causes  $S$ . Plainly, an experiment in which  $B$  is manipulated in this way is a badly designed experiment for the purposes of determining whether  $B$  causes  $S$ . Similarly, an experiment in which the process that changes  $B$  also directly changes  $S$  would be badly designed for this purpose. We need to formulate conditions that restrict the allowable ways of changing  $X$  in an experiment designed to determine whether  $X$  causes  $Y$  in such a way as to rule out possibilities of this sort. Heuristically, we may think of the allowable changes in  $X$  (interventions, as we have been calling them) as processes that satisfy whatever conditions must be met in an ideal experiment designed to determine whether  $X$  causes  $Y$ . There have been a number of attempts to characterize the notion of an intervention more precisely in the recent philosophical literature (e.g., Cartwright and Jones 1991; Spirtes et al. [1993] 2000; Pearl 1995, 2000a; Woodward 1997b; Hausman 1998). The details of my own proposal, which differs in various respects from these other characterizations, are presented in section 3.1. In this section, I focus only on giving the reader intuitive understanding of the notion and its connection to other ideas about causation.

Consider the following experiment. We employ a random number generator, which is assumed to be causally independent of  $A$ , and, depending just on the output of this device, repeatedly physically set the barometer reading by moving the dial to either a high or low reading and fixing it at that position. If it is really true that  $B$  does not cause  $S$ , then we expect that the changes in  $B$  produced by such interventions will no longer be associated with changes in  $S$ . If, on the contrary,  $B$  continues to be correlated with  $S$  under such interventions on  $B$ , this would be strong *prima facie* evidence that  $B$  does cause  $S$ . This example illustrates the idea that interventions involve *exogenous* changes in the variable intervened on. When an intervention occurs on  $B$ , the value of  $B$  is determined entirely by the intervention, in a way that is (causally and probabilistically) independent of the value of  $A$ . In this sense, the intervention “breaks” the previously existing endogenous causal relationship between  $A$  and  $B$ . More generally and slightly more precisely, we may think of an intervention on  $X$  with respect to  $Y$  as an exogenous causal process that changes  $X$  in such a way and under conditions such that if any change occurs in  $Y$ , it occurs only in virtue of  $Y$ 's relationship to  $X$  and not in any other way.

It is important to understand that (i) the information that a variable has been set to some value by an intervention is quite different from (ii) the information that the variable has taken that value as the result of some process that leaves intact the causal structure that has previously generated the values of that variable.<sup>12</sup> Suppose that, in the above example, there is no intervention on  $B$  and that its values continue to be generated by  $A$ . In this case, they will be correlated with the values of  $S$ , which are also generated by  $A$ . If we write  $Pr$  for the probability distribution governing  $A$ ,  $B$ , and  $S$  in the absence of an intervention, then  $Pr(S/B) > Pr(S)$ . By contrast, intervening on  $B$  alters the causal structure of the system in which  $B$  figures, giving  $B$  a new exogenous causal history. This disrupts the previously existing pattern of correlations in the  $A$ - $B$ - $S$  system, leading to a new probability distribution  $Pr_I$  governing those variables. If  $B$  does not cause  $S$ , then we expect that in this new distribution  $Pr_I(S/B) = Pr_I(S)$ . The difference between (i) and (ii) above thus corresponds to the difference between two questions: (1) What would it be reasonable for me to predict regarding the value of  $S$ , if I were to physically manipulate the value of  $B$  in the manner described above? (2) What would it be reasonable to predict regarding the value of  $S$  when I observe the value of  $B$ , assuming that no intervention occurs and whatever system has been generating the values of  $B$  and  $S$  remains intact?

The difference between these two kinds of information (i) and (ii) is often described as the difference between *intervening* and *conditioning* or as the difference between *doing* and *looking*. Pearl (1995) represents the fact that a variable  $X$  has been set to some value by an intervention by means of a new random variable *set*  $X$  and observes that it is not true in general that  $Pr(Y/X) = Pr(Y/set X)$ . (In the above example,  $Pr(S/B) > Pr(S)$ , but  $Pr(S/set B) = Pr(S)$ .) Although this is a useful shorthand I sometimes employ, it is important not to be misled by it.  $X$  and *set*  $X$  are not really different variables, but rather the



same variable embedded in different causal structures that may be associated with different probability distributions.

The idea that interventions represent exogenous changes that alter the causal structure of the system intervened on is closely tied to an idea about how to model the impact of interventions by means of systems of equations that Dan Hausman and I (Hausman and Woodward 1999, forthcoming a) have elsewhere called *modularity*, and which is discussed in more detail in chapter 7. (I claim no originality for this idea, which is closely tied to the econometric idea of autonomy found in Haavelmo and Frisch and to the “wiping out of equations” idea in Strotz and Wold 1960. More recent statements of the idea can be found in Spirtes et al. [1993] 2000 and in Pearl 2000a.<sup>13</sup>) Let us represent the structure of the atmospheric pressure/storm/barometer system by means of the following two equations:

$$(2.3.1) \quad B = aA$$

$$(2.3.2) \quad S = bA$$

We may then represent an intervention on  $B$  by replacing equation (2.3.1) with a different equation (2.3.3),  $B = I$ , specifying that the value of  $B$  is no longer determined by  $S$  but is instead set entirely by the value of the intervention variable  $I$ . If the representation (2.3.1)–(2.3.2) is modular, then it will be possible to carry out this operation of replacing (2.3.1) with (2.3.3) while leaving the other equation in the system (i.e., (2.3.2)) undisturbed. More generally, a system of equations will be modular if it is possible to disrupt or replace (the relationships represented by) any one of the equations in the system by means of an intervention on (the magnitude corresponding to) the dependent variable in that equation, without disrupting any of the other equations.<sup>14</sup> Modularity is thus a feature that a set of *representations* of causal relationships (e.g., equations or a directed graph) may (or may not) possess.

It is natural to suppose that if a system of equations correctly and fully represents the causal structure of some system, then those equations should be modular. One way of motivating this claim appeals to the idea that each equation in the system should represent the operation of a distinct causal mechanism. (Correlatively, each complete set of arrows directed into each variable in a directed graph should also correspond to a distinct mechanism.) If we make the additional plausible assumption that a necessary condition for two mechanisms to be distinct is that it be possible (in principle) to interfere with the operation of one without interfering with the operation of the other and vice versa, we have a justification for requiring that systems of equations that fully and correctly represent causal structure should be modular. (This in turn illustrates how an important causal idea, that of an independent causal mechanism, can be captured in part in manipulationist terms.) For example, in the informal discussion above, we tacitly assumed that the process of intervening to set the barometer reading independently of  $A$  would only disrupt the relationship between  $A$  and  $B$  and would not alter or disrupt the

relationship between  $A$  and  $S$ . Intuitively, the justification for this assumption is that we think that the mechanism or means by which  $A$  affects  $B$  is distinct from the way in which  $A$  affects  $S$ —this is what makes it sensible to think in terms of a hypothetical experiment in which the relationship between  $A$  and  $B$  is disrupted without disrupting the relationship between  $A$  and  $S$ . In what follows, I assume that when causal relationships are correctly and fully represented by systems of equations, each equation will correspond to a distinct causal mechanism and that the equation system will be modular. As we will see, the notion of direct causation and the related notion of a causal route is closely bound up with these ideas.

I turn now to the assessment of **SC** and **NC**. Provided that the notion of an intervention is understood in the appropriate way, I believe that **SC** is extremely plausible. It says, in effect, that if it is possible to manipulate  $Y$  by intervening on  $X$ , then we may conclude that  $X$  causes  $Y$ , regardless of whether the relationship between  $X$  and  $Y$  lacks various other features standardly regarded as necessary for causation: even if  $X$  is not connected to  $Y$  via a spatiotemporally continuous process, even if there is no obvious sense in which there is a transfer of energy from  $X$  to  $Y$ , and so on.

What about **NC**? Consider the causal structure in figure 2.3.1.

Here  $X$  directly causes  $Y$  and  $X$  also directly causes  $Z$ , which in turn directly causes  $Y$ . Assume that all three relationships are linear and can be represented by means of the equations

$$(2.3.4) \quad Y = aX + cZ$$

$$(2.3.5) \quad Z = bX$$

where  $a$ ,  $b$ , and  $c$  are fixed coefficients. Then if  $a = -bc$ , the direct causal influence of  $X$  on  $Y$  will be exactly canceled out by the indirect influence of  $X$  on  $Y$  that is mediated through  $Z$ .<sup>15</sup> Thus, even though  $X$  directly causes and hence (in some relevant sense) causes  $Y$ , there are no manipulations of  $X$  alone that will change  $Y$ . This example involves what Spirtes et al. ([1993] 2000) call a failure of “faithfulness.”

An analogous case involving probabilistic causality is provided by a variant on a well-known example due to Hesslow (1976). Suppose that birth control pills ( $B$ ) directly cause an increased probability of thrombosis ( $T$ ) but also

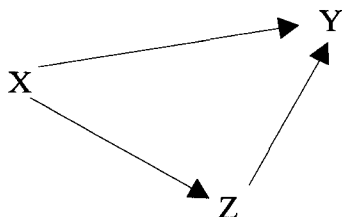


Figure 2.3.1

directly lower the probability of pregnancy ( $P$ ), which is itself a direct positive probabilistic cause of thrombosis. As it happens, the probability increase in  $T$  that is directly due to  $B$  is exactly balanced by the decrease in probability of  $T$  which occurs along the  $B \rightarrow P \rightarrow T$  route, so that the net change in the probability of thrombosis if one takes birth control pills is 0. Thus, there is no possible intervention on  $B$  alone that will change the probability of  $T$ . Nonetheless, it seems clear that there is some sense in which  $B$  is a cause of  $T$  (see figure 2.3.2).

These examples show that we need to distinguish between two notions of “cause.”<sup>16</sup> Let us say that  $X$  is a *total cause* of  $Y$  if and only if it has a non-null total effect on  $Y$ ; that is, if and only if there is some intervention on  $X$  alone (and no other variables) such that for some value of other variables besides  $X$ , this intervention on  $X$  will change the value of  $Y$ . The *total effect* of a change  $\Delta x$  in  $X$  on  $Y$  is then the change in the value of  $Y$  or in the probability distribution of  $Y$  that would result from an intervention on  $X$  alone that changes it by amount  $\Delta x$ , given the values of other variables. (It is assumed that the notion of an intervention is characterized in such a way that the values of other variables that are not descendants of  $X$  are not changed by an intervention on  $X$ ; see section 3.1 below.) For example, in (2.3.4)–(2.3.5) the total effect on  $Y$  of a change of  $\Delta x$  in  $X$  is  $(a + bc)\Delta x$ , and the total effect on  $Y$  of a change  $\Delta z$  in  $Z$  is  $c\Delta z$ . The notion of a total cause contrasts with the notion of a *contributing cause*, which is intended to capture the intuitive idea of  $X$  influencing  $Y$  along some route even if, because of cancellation,  $X$  has no total effect on  $Y$ . The characterization of this notion will occupy us in some detail below but, to anticipate the results of my discussion,  $X$  will be a contributing cause of  $Y$  if and only if it makes a non-null contribution to  $Y$  along some directed path in the sense that for those variables (if any) that are *not* on this path, there is some set of values of those variables such that if the variables were fixed by interventions at those values, there is some intervention on  $X$  that will change the value of  $Y$ . Direct causes will thus always be contributing causes, but there may be contributing causes that are not direct. For example, if we were to add a third equation (2.3.6) to (2.3.4)–(2.3.5) relating an additional variable  $W$  to  $X$  ((2.3.6)  $X = eW$ )—that is, if we were to draw an additional arrow from  $W$  into  $X$ —then, although  $W$  is not a direct cause of  $Y$ ,  $W$  will be a contributing cause of  $Y$  because, freezing the value of  $Z$ , there are interventions on  $W$  that will change  $Y$ .

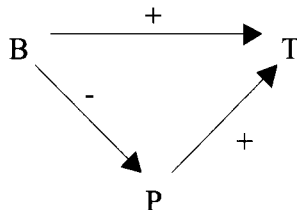


Figure 2.3.2

In the examples in figures 2.3.1 and 2.3.2, the total effect of  $X$  on  $Y$  and of  $B$  on  $T$  is null. In the total or net effect sense of cause,  $X$  is not a cause of  $Y$  and  $B$  is not a cause of  $T$ . Nonetheless, it is also true that  $X$  makes non-null causal contributions to  $Y$  and  $B$  makes a non-null causal contributions to  $T$  along each of two different routes. When we say that  $X$  is a cause of  $Y$  or  $B$  is a cause of  $T$ , it is this second (non-null contribution along a causal route) sense of “cause” that we have in mind. Both directed graphs and equations aim, in the first instance, at the representation of direct rather than total causal relationships. If we have full information about the functional relationships that represent direct causal relationships, we may recover total causal relationships from them, as illustrated above. But directed graphs by themselves do not convey such information. Somewhat more surprisingly, it is also true, as we shall see, that directed graphs do not by themselves convey full information about contributing causal relationships; again, information about functional relationships among direct causes is required.

The notion of a total cause can be easily captured within a manipulationist framework as follows:

**(TC)**  $X$  is a total cause of  $Y$  if and only if there is a possible intervention on  $X$  that will change  $Y$  or the probability distribution of  $Y$ .

In other words, both **SC** and **NC** are defensible if “cause” is interpreted as “total cause.” However, as the above examples illustrate, **NC** is false if “cause” is interpreted to mean “contributing cause.”  $X$  is a contributing cause of  $Y$  in figure 2.3.1, even though there are no possible interventions on  $X$  alone that will change  $Y$ . Similarly for  $B$  and  $T$ . To characterize the notion of a contributing cause (or the notion of a direct cause) in manipulationist terms, we need an analysis that goes beyond **SC** and **NC**.

Providing such an analysis would be unnecessary if all of our purposes would be adequately served just by the “total cause” notion. However, this is not the case. Compare (2.3.4)–(2.3.5) and figure 2.3.1 with equations (2.3.7)–(2.3.8) and figure 2.3.3.

$$(2.3.7) \quad Z = fX$$

$$(2.3.8) \quad Y = gZ$$

Assume now that in (2.3.4)–(2.3.5)  $a \neq -bc$ . Then, in both (2.3.4)–(2.3.5) and (2.3.7)–(2.3.8),  $X$  is a total cause of  $Z$ ,  $Z$  is a total cause of  $Y$ , and  $X$  is a total cause of  $Y$ . According to both sets of equations, an intervention on  $X$  will change  $Z$ , an intervention on  $Z$  will change  $Y$ , and an intervention on  $X$  will change  $Y$ . Nonetheless, (2.3.4)–(2.3.5) and (2.3.7)–(2.3.8) make different claims about the direct causal relationships among  $X$ ,  $Y$ , and  $Z$ . As we will see in more



Figure 2.3.3

detail below, one way this difference manifests itself is in the different predictions (2.3.4)–(2.3.5) and (2.3.7)–(2.3.8) make about what will happen under *combinations* of interventions, rather than just single interventions: (2.3.4)–(2.3.5) predict that if we were to change the value of  $X$  while at the same time we held fixed the value of  $Z$  in a way that was independent of the change in the value of  $X$ ,  $Y$  would change. By contrast, (2.3.7)–(2.3.8) predict that  $Y$  would not change under this combination of interventions. Thus, one reason we need the notion of a direct cause (and the associated notion of a causal path or route) is to capture or represent facts about what will happen under combinations of interventions; such facts are not captured just by information about total causal relationships and about what will happen under single interventions.

A second reason we need the notion of a direct cause, also discussed in more detail below, is that this notion is needed to formulate plausible claims connecting causes to probabilities: in explaining what needs to be controlled for or conditioned on when we test for whether  $X$  causes  $Y$ , we need information about direct and not just total causal relationships. For closely related reasons, we also need information about direct causal relationships to characterize the notion of an intervention.

A third reason we need the notion of direct causation is to capture the ideas about distinctness of causal mechanisms sketched above. Return to equations (2.3.4)–(2.3.5) and substitute  $bX$  for  $Z$  obtaining (2.3.9)  $Y = hX$  where  $h = a + bc$ . (2.3.9) correctly describes how  $Y$  will change under an intervention on  $X$ . What, if anything, would be wrong with using (2.3.9) alone rather than (2.3.4)–(2.3.5) to represent the system that we are trying to model? If (2.3.4)–(2.3.5) correctly represent this system, then it consists of two distinct mechanisms: a mechanism represented by (2.3.5), by which  $X$  influences  $Z$ , and an independent mechanism represented by (2.3.4) by which  $X$  and  $Z$  together influence  $Y$ . If we use only equation (2.3.9), we misrepresent (or at least fail to represent) this fact. In effect, (2.3.9) collapses two distinct mechanisms by which  $X$  influences  $Y$  into just one. Similarly, figure 2.3.2 tells us that there are two distinct mechanisms connecting  $B$ ,  $T$ , and  $P$  (a mechanism in which  $B$  and  $P$  influence  $T$  and a second, independent mechanism by which  $B$  influences  $P$ ), or alternatively, two distinct routes or paths by which  $B$  influences  $T$  (a direct route along which  $B$  makes a positive contribution to  $T$  and an indirect route along which  $B$  has a negative effect on  $T$  in virtue of lowering the probability of  $P$ ). We could, of course, write the probability of  $T$  as a function of  $B$  alone, but again, this would be to collapse two distinct mechanisms or two distinct pathways by which  $B$  influences the probability of  $T$  into one.

As suggested above, if it is really true that the mechanism by which  $B$  influences  $P$  is distinct from the mechanism by which  $B$  and  $P$  influence  $T$ , then it should be possible to interfere with one of these without interfering with the other. We might simulate the effects of such an interference by, for example, either (a) comparing the incidence of thrombosis among those who take birth control pills with those who do not in a population of women all of whom are infertile or who also employ other methods of contraception and hence cannot become pregnant, or (b) comparing the incidence of thrombosis

among pregnant women who continue to take birth control pills with the incidence of thrombosis among pregnant women who do not take birth control pills. In such populations the causal relationship or mechanism ( $B \rightarrow P$ ) connecting  $B$  to  $P$  is disrupted, because whether or not these women become pregnant is not influenced by whether they take birth control pills. If the claim about distinctness of mechanisms is correct, then under some disruptions of the  $B \rightarrow P$  mechanism (including, plausibly, the ones just described), the mechanism linking  $B$  and  $P$  to  $T$  should continue to operate as before, and this in turn means that the incidence of thrombosis should be higher among pill takers in one or both populations<sup>17</sup> (i.e., that the positive contribution of  $B$  to  $T$  represented by the direct effect of  $B$  on  $T$  should be preserved). In other words, if figure 2.3.2 is correct, we should expect such comparisons to reveal the positive causal contribution of  $B$  to  $T$  that is masked by the counterbalancing negative effect of  $B$  on  $T$  that operates through  $P$ . Again, we need the notion of direct causation to capture these ideas.

I conclude from this that the notion of a direct cause is *a* (but not the only) legitimate notion of cause and that not all of the work done by this notion can be accomplished just with the notion of a total cause. As suggested above, this assumption is implicit in the use of directed graphs and systems of equations to represent causal relationships and also follows from the underlying logic of a manipulability approach to causation. Even when  $a = -bc$  in (2.3.4)–(2.3.5), one may still use  $X$  to change or manipulate  $Y$ —all that one has to do is to fix  $Z$  at some value and then wiggle  $X$ . Thus, there is a perfectly good sense in which  $X$  remains a means to changing  $Y$ , and this, I claim, is enough to establish that there is a legitimate sense in which  $X$  is a cause of  $Y$ . That sense is captured by the notion of a direct (or contributing) cause.

In what follows, I explore the possibility of capturing the notion of a contributing cause within the framework of a manipulability account. My strategy is to formulate first a necessary and sufficient condition for  $X$  to be a direct cause of  $Y$  and then to use this formulation to arrive at a necessary and sufficient condition for  $X$  to be a contributing cause of  $Y$ .

Pursuing this strategy in connection with the causal structure described in figure 2.3.1, consider in more detail the suggestion adumbrated above. If the causal structure in figure 2.3.1 is correct, then there should be some intervention that changes the value of  $X$  which would change the value of  $Y$  in the way indicated by figure 2.3.1 (and equations (2.3.4)–(2.3.5)) when the value of  $Z$  is held fixed at some value by an additional intervention. Intervening to hold  $Z$  fixed at some value by an intervention while changing the value of  $X$  by an intervention requires that  $Z$  is set in such a way that its value is entirely dictated by the intervention, rather than by the value of  $X$ . At the same time, the value of  $X$  is changed by some other process, also satisfying the conditions for an intervention, that is causally independent of and uncorrelated with the process that fixes the value of  $Z$ . (I emphasize again that all that is required is that there be *some* value of  $Z$  for which the condition that *some* intervention on  $X$  will change  $Y$  is satisfied; it is not required that the condition hold for all values of  $Z$  or that all interventions on  $X$  change  $Y$  for some value of  $Z$ .)

If there are no possible interventions on  $X$  that, holding  $Z$  fixed at some value, change the value of  $Y$ , it follows that  $X$  does not directly cause  $Y$  and hence that figure 2.3.1 is not the correct causal structure. Similarly, as suggested above, if it is true that  $B$  is a direct cause of  $T$ , then, if we were to intervene to hold fixed  $P$  (an operation that, as noted above, we might simulate by considering a group of women all of whom are already pregnant or a group of women who cannot become pregnant for physiological reasons that are independent of whether they take birth control pills) and manipulate the value of  $B$ , we would expect the probability of thrombosis to change for at least one value of  $P$ .<sup>18</sup>

This gives us a set of claims about what would happen under hypothetical interventions that are plausible necessary conditions for it to be true that  $X$  is a direct cause of  $Y$  in structures like figure 2.3.1. To what extent do these claims generalize; that is, to what extent are they also necessary conditions for direct causation in other causal structures? Consider again the alternative causal structure (figure 2.3.3) in which there is a single route from  $X$  to  $Y$  that goes through  $Z$ . In (2.3.3),  $X$  is not a direct cause of  $Y$ . In such structures, it is still a necessary condition for  $X$  to be a direct cause of  $Y$  that there be a possible intervention on  $X$  that changes  $Y$  when  $Z$  is fixed at some value. It is because it is true that if we were to fix  $Z$  at some value and then intervene to change  $X$ , we would observe no change in  $Y$  that  $X$  fails to qualify as a direct cause of  $Y$ .

Next, consider a structure (figure 2.3.4) in which there are two indirect routes from  $X$  to  $Y$ , one through  $Z$  and one through  $W$ , and for which we want to know whether there is also a direct causal route from  $X$  to  $Y$ . Here we face the possibility that if there is such a route, the influence of  $X$  on  $Y$  along this route may be canceled by the combined effect of  $X$  on  $Y$  along both of the indirect routes from  $X$  to  $Y$ . In this structure, a (nontrivial) necessary condition for  $X$  to directly cause  $Y$  is that  $Y$  changes under some intervention that changes  $X$  when both  $Z$  and  $W$  are held fixed at some value.

Is there some way of generalizing these observations to capture what it is for  $X$  to be a direct cause of  $Y$  in all causal structures? We may do so as follows. Suppose that there are  $n$  directed paths or routes from  $X$  to  $Y$  that contain intermediate causes, with  $Z_i$  an intermediate cause on the  $i$ th route. Then

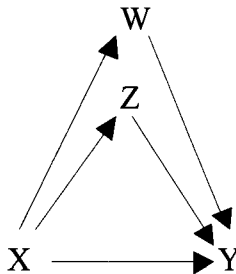


Figure 2.3.4

a necessary and sufficient condition for it to be true that  $X$  is a direct cause of  $Y$  is that there be a possible intervention on  $X$  that changes  $Y$  when each of the intermediate causes along all of the routes from  $X$  to  $Y$  that go through  $Z_1, Z_2 \dots Z_n$  is held fixed at some value by interventions. Moreover, it should also be clear that if there are other intermediate causes  $W_1 \dots W_m$  along these alternative routes in addition to  $Z_1 \dots Z_n$ , it will make no difference to our assessment of whether  $X$  is a direct cause of  $Y$  if we hold these fixed as well; that is, for the purpose of defining whether  $X$  is a direct cause of  $Y$ , it will make no difference if we ask what will happen to  $Y$  if we manipulate  $X$  while holding only  $Z_1, Z_2 \dots Z_n$  fixed or if we hold both  $Z_1 \dots Z_n$  fixed and one or more of  $W_1 \dots W_m$  fixed as well while intervening on  $X$ . In other words, what is crucial if we want to determine whether  $X$  is a direct cause of  $Y$  is that we hold fixed by interventions at some value at least one intermediate cause along each alternative route, but it does not matter if we also hold fixed more than one such intermediate cause. In addition, if  $X$  is a direct cause of  $Y$ , then some intervention on  $X$  should change  $Y$  even if we not only hold fixed all intermediate causes  $Z_i$  between  $X$  and  $Y$  but also hold fixed any other variable  $V_i$  besides  $X, Y$ , and  $Z_i$  where  $V_i$  is some causal ancestor of  $X$  or some causal descendant of  $Y$  or where  $V_i$  is neither an ancestor nor a descendant of  $X$  and  $Y$ . In short, as long as we hold fixed some variable along each of the alternative causal routes from  $X$  to  $Y$  besides the direct route, it does not matter what else we hold fixed. In particular, if we hold fixed *all* other variables in the variable set  $\mathbf{V}$  that describes the system of interest and find that there is a possible manipulation of  $X$  that changes  $Y$ , it follows that  $X$  is a direct cause of  $Y$  with respect to that variable set. Conversely, if  $X$  is a direct cause of  $Y$ , then even holding fixed all of the other variables in the system of interest, there must be some intervention on  $X$  that changes  $Y$  (or, in the case in which  $X$  is a probabilistic cause, the probability distribution of  $Y$ ) for some value of those other variables.

Putting these observations together, we may define the notion of a direct cause as follows<sup>19</sup>:

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

Obviously, this definition of what it is for  $X$  to be a direct cause of  $Y$  relativizes the notion to a set of additional variables  $\mathbf{V}$ .  $X$  may be a direct cause of  $Y$  with respect to variables in the set  $\mathbf{V}^*$  but not with respect to a different set of variables  $\mathbf{V}^{**}$ . In some respects, this relativization seems undisturbing; it merely reflects the fact that any description of causal relationships reflects a choice of level of analysis. Consider a concrete example. With respect to a set that includes variables like  $A$ 's desire for revenge,  $A$ 's pulling the trigger of the gun,  $A$ 's hitting  $B$  on the head with a rock,  $A$ 's poisoning  $B$ 's drink, and  $B$ 's death,  $A$ 's desire for revenge may be a direct cause of his pulling the trigger, which may in turn be a direct cause of  $B$ 's death. If we consider instead a



different and more fine-grained set of variables, this may not be true. For example, suppose that pulling the trigger causes the release of a spring, which causes a hammer to strike a cartridge, which causes the cartridge to explode, which propels the bullet out of the barrel of the gun along a trajectory such that it strikes *B*'s heart. With respect to these additional variables, it will not be true that the trigger pulling is a direct cause of *B*'s death. However, at least in this sort of case, once the set of variables  $\mathbf{V}$  is specified, it is a fully objective matter, dependent on the causal structure of the phenomena we are attempting to capture, whether *X* is a direct cause of *Y* with respect to  $\mathbf{V}$ . The sort of relativization illustrated by this example does not introduce an undesirable kind of subjectivism into the characterization of "direct causation" in the sense that it makes whether or not *X* is a direct cause of *Y* dependent on the beliefs or psychological state of human investigators. Moreover, given the way the notion of a contributing cause will be defined below, as long as there is a single causal route from *X* to *Y*, if *X* is a contributing cause of *Y*, *X* will remain a cause of *Y* (although not a direct cause of *Y*) if additional variables *Z* are interpolated between *X* and *Y* along this route. Thus, in the example above, if *A*'s pulling the trigger is a contributing cause of *B*'s death with respect to a variable set that does not include the release of the spring, the hammer striking the cartridge, and so on, it remains a contributing cause when the variable set is expanded to include these variables.

There are, however, other consequences of the relativization to a variable set in the definition **DC** that may seem more problematic. Consider again the structure in figure 2.3.1, with exact cancellation along the two routes. With respect to a variable set that includes *Z*, *X* is a direct cause of *Y* and hence a (contributing) cause of *Y*. However, with respect to a variable set that does not include *Z*, *X* will not be a direct cause of *Y*. Moreover, because of the way the notion of direct causation figures in the definition of (contributing) "cause" adopted below, it also follows that *X* is not a contributing cause of *Y*. Thus, whether *X* counts as a contributing cause of *Y* (and not just a direct cause) depends on whether *Z* is included in the variable set.<sup>20</sup>

This feature of **DC** illustrates a more general fact about the use of systems of equations and directed graphs to represent contributing causal relationships: which relationships are represented as causal in the contributing sense (and not just directly causal) is sensitive to which variables are employed in the representation. This raises issues that I address in more detail below, but several brief remarks may be helpful at this point. First, in my view, the choice of variables in a representation reflects those possibilities that we are willing to "take seriously" in a sense to be described in section 2.8. For reasons that will emerge in that section, I believe that some relativization to "serious possibilities" will be a feature of any plausible theory of causation. In this sense, representation sensitivity is not a feature that is peculiar to **DC** or to the manipulability theory.

Second, it is worth noting that whether or not a variable *Z* is included in the variable set that is held fixed will not affect the evaluation of counterfactuals of the following form: if an intervention were to occur on *X* (but not

on  $Z$ ), then  $Y$  would change in such and such a way. That is, given a structure like that in figure 2.3.1, an investigator who includes  $Z$  in his variable set and an investigator who does not will agree about what will happen to  $Y$  under an intervention just on  $X$  and no other variables; for example, in the case in which there is cancellation along the direct and indirect routes, both will agree that an intervention on  $X$  will produce no change in  $Y$ . Although in this case both investigators will agree about the total causal relationship between  $X$  and  $Y$ , we shall see below that there are other respects in which total causal relationships, as well as direct causal relationships, are plausibly regarded as representation-dependent. This suggests that there is a sense in which facts about patterns of counterfactual dependence are more basic than facts about what causes what in *either* the total cause or direct cause sense—a theme to which I will return below.

Given the definition **DC** of what it is for  $X$  to be a direct cause of  $Y$ , we may then formulate a necessary condition, expressed in terms of claims about the outcomes of hypothetical interventions, for  $X$  to be a contributing, (type-level) cause of  $Y$  as follows:

**(NC\*)** If  $X$  is a contributing type-level cause of  $Y$  with respect to the variable set  $\mathbf{V}$ , then there is a directed path from  $X$  to  $Y$  such that each link in this path is a direct causal relationship; that is, there are intermediate variables along this path,  $Z_1 \dots Z_n$ , such that  $X$  is a direct cause of  $Z_1$ , which is a direct cause of  $Z_2$ , which is a direct cause of  $\dots Z_n$ , which is a direct cause of  $Y$ . Put differently, if  $X$  causes  $Y$ , then  $X$  must either be a direct cause of  $Y$  or there must be a causal chain, each link of which involves a relationship of direct causation, extending from  $X$  to  $Y$ .

**NC\*** does *not* require the assumption that contributing causation is transitive. Nor does it require related transitivity-like assumptions connecting direct causation to contributing causation (e.g., if  $X$  is a direct cause of  $Z$  and  $Z$  a contributing cause of  $Y$ , then  $X$  is a contributing cause of  $Y$ ). Assumptions about transitivity involve sufficient conditions for causation, and **(NC\*)** purports to provide only a necessary condition. The proper role for transitivity and related requirements, if they are assumed at all, is in the statement of a sufficient condition for causation—the topic to which I now turn.

If it were justifiable to make the transitivity assumptions described above, we could simply replace the “if”s in **NC\*** with “if and only if”s and we would have a sufficient as well as a necessary condition for contributing causation. In fact, however, such transitivity assumptions are dubious. Consider the following example, which is due to McDermott (1995 p. 531).<sup>21</sup> A dog bites off my right forefinger. The next day I detonate a bomb by using my left forefinger. If I had not lost my right finger, I would have used it instead to detonate the bomb. The bite causes me to use my left finger, which causes the bomb to explode, but (it seems) the bite does not cause the bomb to explode. Although McDermott’s version of the example involves token causation, if we think of the setup he describes as a potentially repeatable process involving type-level variables (with  $B$  specifying whether or not a bite occurs,  $L$  taking one of *three*

values, according to whether the left finger is used, the right finger is used, or the detonating button is not pushed at all, and  $E$  specifying whether or not the bomb explodes), then the example is translatable in an obvious way into a potential counterexample to the transitivity of type causation. It is also a counterexample to the proposal that  $\text{NC}^*$  may be turned into a sufficient condition for causation by replacing references to necessary conditions with references to sufficient conditions:  $B$  is a direct cause of  $L$ , which is in turn a direct cause of  $E$ , so that there is a directed path from  $B$  to  $E$ , but  $B$  does not cause  $E$ .<sup>22</sup>

Fortunately, a natural treatment of this case is available within a manipulability framework.<sup>23</sup> In the structure under discussion, an intervention that changes whether or not a bite occurs changes whether I use my left or right finger to detonate the bomb. Moreover, an intervention that changes the situation from one in which I use my left finger to detonate the bomb to a situation in which I do not detonate the bomb at all (with either finger) changes whether the bomb explodes. However, changing whether I use my left or right finger to detonate the bomb does not change whether the bomb explodes. In other words, although there are changes in the value of  $L$  (whether I use my right or left finger) that are sensitive to whether or not an intervention occurs that changes the value of  $B$ , and *other* changes in the value of  $L$  (whether I use my left finger rather than not pressing the detonating button at all) to which the value of  $E$  is sensitive, there is no set of changes in the value of  $L$  that fulfills both these roles. The changes in the value of  $L$  on which the value of  $E$  depends are completely different from the changes in the value of  $L$  that are influenced by the value of  $B$ , so that there is no overall sensitivity of the value of  $E$  to the value of  $B$  along the only route connecting  $B$  to  $E$ : manipulating the value of  $B$  does not change the value of  $E$ . I believe that it is this fact (of insensitivity along the only route connecting  $B$  to  $E$ ) that makes us judge that  $B$  does not cause  $E$  and that transitivity fails.<sup>24</sup>

Formally, the case is one in which  $Z = F(X)$  and  $F$  maps all values of  $X$  into  $z_1$  except for the value  $x_2$ , of  $X$  which is mapped into  $F(x_2) = z_2 \neq z_1$ , and  $Y = G(Z)$  with  $G$  mapping both  $z_1$  and  $z_2$  into  $y_1$  but taking some third value of  $Z$ ,  $z_3$  ( $z_3 \neq z_1$  and  $z_3 \neq z_2$ ) into  $y_2 \neq y_1$ . Because there is some manipulation of the value of  $X$  that will change the value of  $Z$  (even when the value of  $Y$  is fixed), we find it natural to conclude that  $X$  is both a total cause and a direct cause of  $Z$ , as **SC** and **DC** require. Moreover, because there is some manipulation of the value of  $Z$  that will change the value of  $Y$ , even when the value of  $X$  is held fixed, we also conclude that  $Z$  is both a total and a direct cause of  $Y$ . However, because there is no change in  $X$  that will change the value of  $Y$  along the single route from  $X$  to  $Y$  and because the case is not one in which there are multiple routes between  $X$  and  $Y$  along which cancellation might occur, we find it natural to conclude that  $X$  does not in any sense (either total or contributing) cause  $Y$ . Transitivity fails because the functions  $F$  and  $G$  compose in such a way that along the only route from  $X$  to  $Y$ , the composite function  $Y = G(F(X))$  is such that the value of  $Y$  is not sensitive at all to changes in the value of  $X$ . Note also that this sort of failure of transitivity is quite different from the cancellation of causal contributions along different routes represented by

figure 2.3.1 with  $a = -bc$ . In the present case, transitivity fails because the relevant functions compose along a single route in such a way that  $Y$  is not sensitive to changes in  $X$ . In the figure 2.3.1 case, there is no failure of sensitivity due to the way functions compose along either of the two routes from  $X$  to  $Y$ .<sup>25</sup> Within a manipulability framework, this is an important difference. In figure 2.3.1, intervening on  $X$  (while holding  $Z$  fixed at some value) allows us to change the value of  $Y$ . Hence, there is an obvious motivation for regarding  $X$  as a cause of  $Y$ . By contrast, in the dog bite example, no change in the value of  $X(B)$ , even holding some third variable such as  $Z(L)$  fixed, will change the value of  $Y(E)$ .

If this analysis is correct, then one natural way to deal with failures of transitivity of the sort under discussion is to require that for  $X$  to be a contributing cause of  $Y$ , not only must there be at least one chain of direct causal relationships (a directed path or route) from  $X$  to  $Y$ , but it must also be the case that the value of  $Y$  is sensitive along that path to some interventions that change the value of  $X$ ; that is, it must be the case that there is a directed path from  $X$  to  $Y$  such that an intervention on  $X$  will change  $Y$  when all variables that are not on this path, including intermediate variables on *other* paths between  $X$  and  $Y$  and variables along any other paths leading into  $Y$ , are fixed at some value.<sup>26</sup>

Putting this idea together with **NC\*** suggests the following necessary and sufficient conditions for  $X$  to be a (type-level) direct cause of  $Y$  and for  $X$  to a (type-level) contributing cause of  $Y$ . I label these **M** for “manipulability theory”:

**(M)** A necessary and sufficient condition for  $X$  to be a (type-level) direct cause of  $Y$  with respect to a variable set  $\mathbf{V}$  is that there be a possible intervention on  $X$  that will change  $Y$  or the probability distribution of  $Y$  when one holds fixed at some value all other variables  $Z_i$  in  $\mathbf{V}$ . A necessary and sufficient condition for  $X$  to be a (type-level) *contributing cause* of  $Y$  with respect to variable set  $\mathbf{V}$  is that (i) there be a directed path from  $X$  to  $Y$  such that each link in this path is a direct causal relationship; that is, a set of variables  $Z_1 \dots Z_n$  such that  $X$  is a direct cause of  $Z_1$ , which is in turn a direct cause of  $Z_2$ , which is a direct cause of  $\dots Z_n$ , which is a direct cause of  $Y$ , and that (ii) there be some intervention on  $X$  that will change  $Y$  when all other variables in  $\mathbf{V}$  that are not on this path are fixed at some value. If there is only one path  $P$  from  $X$  to  $Y$  or if the only alternative path from  $X$  to  $Y$  besides  $P$  contains no intermediate variables (i.e., is direct), then  $X$  is a contributing cause of  $Y$  as long as there is some intervention on  $X$  that will change the value of  $Y$ , for some values of the other variables in  $\mathbf{V}$ .

To determine whether **M** is satisfied, it is not really necessary to fix all off-path variables at all possible combinations of values. Instead, we may proceed as follows. We first draw a causal graph that represents all of the direct causal relationships between  $X$ ,  $Y$ , and all other variables in  $\mathbf{V}$ . In constructing the graph, we draw an arrow from one variable  $U$  to another  $V$  if and only if  $U$  is a direct cause of  $V$  in the sense specified in **DC**. We then check all of the routes or directed paths from  $X$  to  $Y$ . For each route  $P_i$  from  $X$  to  $Y$ , we freeze at least one intermediate variable at each of its various possible values along all *other*

routes from  $X$  to  $Y$  containing intermediate variables. We also fix at each of its possible values all other direct causes of  $Y$  that are not on any directed path from  $X$  to  $Y$ . If for some combination of values of these off-path variables, some intervention on  $X$  will change the value of  $Y$ , then  $X$  is a contributing cause of  $Y$ . Alternatively, we check to see whether there is some possible combination of values of all of the direct causes of  $Y$  that are not on the path  $P_i$  such that with these values fixed, some intervention on  $X$  will change the value of  $Y$ . We repeat this process for each path from  $X$  to  $Y$ . As long as there is at least one path such that when the appropriate variables not on this path are fixed at some value, changing  $X$  changes  $Y$ , then  $X$  is a contributing cause of  $Y$ . If there is only one directed path from  $X$  to  $Y$  and the only direct cause of  $Y$  is on this path (as in the dog bite example), then  $X$  is a contributing cause of  $Y$  if and only if some intervention on  $X$  changes  $Y$ . If there is a path from  $X$  to  $Y$  containing an intermediate variable and a second (direct) path from  $X$  to  $Y$  that contains no intermediate variables, then there is no operation of freezing intermediate variables along the second (direct) path, and  $X$  is a contributing cause of  $Y$  if and only if some intervention on  $X$  changes  $Y$ , for some combination of values of the other off-path variables in  $\mathbf{V}$ . If there are no routes from  $X$  to  $Y$  such that changes in  $X$  change  $Y$ , when intermediate variables along other routes and other off-path variables are fixed at some value, then  $X$  is not a contributing cause of  $Y$ . For example, in figure 2.3.1 and equations (2.3.4)–(2.3.5), there are two directed paths from  $X$  to  $Y$ , and the only remaining variable in  $\mathbf{V}$ , which is  $Z$ , is on one of these paths. To evaluate whether  $X$  is a contributing cause of  $Y$ , we first consider the direct route from  $X$  to  $Y$  and check whether, when the intermediate variable  $Z$  on the other route from  $X$  to  $Y$  is fixed at some value, there is an intervention on  $X$  that will change  $Y$ . The answer to this question is yes, and hence  $X$  is a contributing cause of  $Y$ .

Although **M** looks complex, the underlying idea is quite simple and intuitive:  $X$  is a contributing cause of  $Y$  with respect to  $\mathbf{V}$  if and only if there are changes in  $X$  that will change  $Y$  when the right other variables are held fixed at some value. The right other variables are defined with respect to the directed paths into  $Y$ : for each directed path between  $X$  and  $Y$  we are to hold fixed at each possible value at least one variable that is not on that path (or equivalently, all direct causes of  $Y$  that are not on that path) and then determine whether there is a change in  $X$  that will change  $Y$ .

**M** formulates conditions having to do with what would happen under hypothetical interventions that are both necessary and sufficient for  $X$  to be a contributing (type) cause of  $Y$ . Similarly, **TC** formulates conditions about what would happen under interventions that are necessary and sufficient for  $X$  to be a total (type) cause of  $Y$ . In my view, this is what we should demand of a satisfactory manipulability theory. If there are facts about what would happen to  $Y$  under hypothetical interventions on  $X$  that are sufficient for  $X$  to cause  $Y$  but there are no such facts that are necessary for  $X$  to cause  $Y$ , we would then face the possibility that there is some other set of conditions, having nothing to do with facts about what would happen under manipulation of  $X$ , that are also

sufficient for  $X$  to cause  $Y$  and puzzling questions about the relationship between these two sets of conditions and why they are both relevant to causation. Similarly, if there are facts about what would happen under manipulations that were necessary for causation but no such facts that are sufficient, then a manipulability account of causation again would be incomplete: there would be some additional content to causal claims that could not be cashed out in terms of facts about what would happen under hypothetical manipulations. By providing both necessary and sufficient conditions for causation, **M** and **TC** give us a way of fully capturing or cashing out the content of causal claims in terms of facts about what would happen under interventions.

On the conception of causation embodied in **M**, we may think of each system of causal relationships as a codification of a distinctive set of claims about what will happen under various combinations of hypothetical interventions or manipulations, distinctive in the sense that any other alternative causal structure will imply some different claims about what will happen under such interventions. For example, the alternative causal structures in figures 2.3.1 and 2.3.3 make different predictions about what will happen to  $Y$  if we fix  $Z$  and appropriately intervene on  $X$ : figure 2.3.1 claims that when  $Z$  is fixed at some value, there is an intervention on  $X$  that will change  $Y$ ; whereas figure 2.3.3 denies this. Conversely, each completely specified set of claims about what will happen to each of the various variables in some set under various possible manipulations of each of the other variables, singly and in combination, will correspond to a distinct causal structure in the sense that each such structure will make distinctive claims about direct causal relationships. In other words, once we fix our representational repertory (i.e., once we choose a set of variables to represent the quantities whose causal relationships we are interested in assessing), then two theories will make different claims about causal relationships among these variables if and only if they make different claims about what will happen under some combination of interventions.<sup>27</sup> Putting this in the form of a slogan, we can say that manipulability accounts are committed to the following: *No causal difference without a difference in manipulability relations, and no difference in manipulability relations without a causal difference.*

## 2.4 Causes and Probabilities

As explained above, one reason we need the notion of a direct cause is to capture or represent facts about what will happen under combinations of interventions. This section argues that we also need the notion of a direct cause for another reason: to formulate plausible conditions connecting causal claims to claims about conditional probabilities.

I begin by focusing on one of the best-known proposals about this connection, the condition (**CC**) formulated by Nancy Cartwright (1983b, p. 26). (Broadly similar proposals are endorsed by a number of other writers, including

Eells 1991 and Eells and Sober 1983.) This will also give us a chance to compare **(CC)** with **(M)** and **(TC)**. According to **(CC)**:

*C* causes *E* if and only if *C* increases the probability of *E* in every situation that is otherwise causally homogeneous with respect to *E*.

Following Cartwright, let us define a complete set of causal factors for *E* as the set of all factors  $C_i$  such that either  $C_i$  causes *E* or  $C_i$  causes not *E*. If there are  $n$  such factors, then there will be  $2^n$  different state descriptions  $K_j = \bigwedge \pm C_i$  corresponding to each possible conjunction of causal factors. Using this apparatus, **CC** can be stated more precisely as follows:

*C* causes *E* iff  $Pr(E/C \cdot K_j) > Pr(E/K_j)$  for all state descriptions  $K_j$  over the set  $\{C_i\}$  where  $\{C_i\}$  satisfies

- (i) If  $C_i$  is in  $\{C_i\}$ , then  $C_i$  causes either *E* or not *E*.
- (ii) *C* is not in  $\{C_i\}$ .
- (iii) For all  $D$ , if  $D$  causes *E* or  $D$  causes not *E*, then either  $D = C$  or  $D$  is in  $\{C_i\}$ .
- (iv) If  $C_i$  is in  $\{C_i\}$ , then *C* does not cause  $C_i$ .

**CC** can be interpreted in at least four ways: “cause” on its left hand side (l.h.s.) can be interpreted as either “total cause” or as “contributing cause,” and similarly, the conjunctions of “other causes” of *E* on which we are told to condition on the right hand side (r.h.s.) of **CC** may be understood to mean either total or contributing causes. I will argue that if anything along the lines of **CC** is to be remotely plausible, whether as an account of contributing or total causation, “cause” on the r.h.s. of **CC** (i.e., in (i)–(iv)) cannot mean “total cause”; instead, **CC** must be understood in such a way as to take account of direct causal relationships.

In addition to the fact that it is unclear whether **CC** refers to total or contributing causal relationships, **CC** differs from the characterizations of causation embodied in **TC** and **M** in a number of other respects. First, and most obvious, **CC** is formulated in terms of conditional probability relationships rather than in terms of counterfactuals about what will happen under interventions. Second, **CC** is formulated in terms of dichotomous variables corresponding to the presence or absence of various causal factors and requires that causes (within our framework, a change in the value of the cause variable from absent to present) raise the probabilities of their effect. By contrast, both the contributing and total notions of cause described above are defined for nondichotomous as well as dichotomous variables. In addition, **TC** and **M** are not formulated in terms of probability raising, but require instead that a change in the value of the cause variable change the value or the probability distribution of the effect variable. This last difference is in part terminological. **CC** attempts to capture the notion of a positive causal factor or of a promoting cause as opposed to the notion of a negative causal factor or a preventive or inhibiting cause. By contrast, as explained above, both **TC** and **M** attempt to capture the broader notion of one variable being causally relevant (either positively or negatively) to another. If our interest is in formulating

a general connection between causation and probability, a variety of considerations seem to me favor this broader usage.<sup>28</sup>

If we look for connections between causation and facts about probabilities that are in the spirit of **CC** but cover nondichotomous variables and capture this broader notion of causal relevance, they presumably will be proposals of the following form:  $X$  causes  $Y$  if and only if  $X$  and  $Y$  are dependent conditional on certain other factors  $F$ , where (one assumes) the specification of these additional factors will depend on whether “cause” on the l.h.s. of this biconditional is interpreted as “total cause” or “contributing cause.” In other words, the question is this: What should be held fixed (i.e., conditioned on) if the conditional dependence of  $C$  on  $E$  is to be used as a test for whether  $C$  causes  $E$  for various interpretations of “cause”?

Let us first explore this question when “causes” means “is a contributing cause.” **CC** says that the other factors  $F$  that should be conditioned on are all other “causes” of  $E$ , with the exception of those causes of  $E$  that are on a causal path or route from  $C$  to  $E$ . (As Cartwright 1983b, p. 26 explains, condition (iv) in **CC** is intended to exclude conditioning on such factors.) The motivation for not holding fixed causal factors that are between  $C$  and  $E$  may seem obvious. If we are dealing with a causal structure like that represented in figure 2.3.3 in which there is a single directed path from  $X$  to  $Y$  with  $Z$  as a causally intermediate variable, then one would expect that conditional on  $Z$ ,  $X$ , and  $Y$  will be independent. Hence, if  $Z$  is one of the background factors on which we condition when we test for whether  $X$  causes  $Y$ , we will reach the mistaken conclusion that  $X$  does not cause  $Y$ . However, as Cartwright herself recognizes (1983b, p. 30, 1989b, pp. 95ff), the claim that, as (iv) requires, we should *never* control for such intermediate variables is too strong.<sup>29</sup> Suppose that we are presented with a triangular structure like that in (2.3.4)–(2.3.5) and figure 2.3.1, in which both  $X$  and  $Z$  are direct causes of  $Y$ , and  $X$  is also a direct cause of  $Z$ . Clearly, if the direct causal connection between  $X$  and  $Y$  is to reveal itself in the probabilistic dependence of  $Y$  on  $X$  conditional on some appropriately chosen set of other factors, these other factors must somehow include an adjustment for the role of  $Z$  that is causally intermediate between  $X$  and  $Y$ . That is, to capture the direct influence of  $X$  on  $Y$ , we must in some way control for or correct for the influence of  $Z$  on  $Y$ . Moreover, because, as we have seen, the total cause structure is the same in both (2.3.4)–(2.3.5) and (2.3.7)–(2.3.8), we need to know the direct causal relationships (and not just the total causal relationships) among  $X$ ,  $Y$ , and  $Z$  to know what to control for when we test for, for example, whether  $X$  is a (contributing) cause of  $Y$ . One way of seeing this is to note that failure to control for  $Z$  in the case in which there is exact cancellation along the two routes will lead to the mistaken conclusion that  $X$  is not a (contributing) cause of  $Y$ . Examples of this sort show that, contrary to what is commonly assumed in the philosophical literature, there is a notion of causation, the notion of  $X$ 's being a contributing cause of  $Y$ , for which it is mistaken to impose the requirement that nothing causally intermediate between  $X$  and  $Y$  (or nothing causally downstream from  $X$ ) be held fixed.



The solution to the problem of what should be “held fixed” in determining whether  $X$  is a contributing cause of  $Y$  is given by **M**, which tells us to assess whether changes in  $X$  influence  $Y$  along a particular path or route, holding fixed by interventions any intermediate variables along other routes. But to employ this solution, one must make use of the notion of a “direct cause,” for it is this notion that is used to define the notion of a causal path. More generally, in determining which variables to control for to determine whether  $X$  is a contributing cause of  $Y$ , it is not enough to consider only which other variables are contributing (or total) causes of  $Y$ . It matters how in detail those other variables are connected; that is, what the full set of direct causal relations and causal routes are. It is just this information about direct causal relationships that is contained in the associated equational or directed graph structure, and this in turn suggests that information about such structures is essential if the sort of project represented by **CC** (the project of formulating systematic relationships between causal claims and conditional probability relationships) is to have any hope of success.

This point of view (that what we need to “control for” will depend on direct causal relationships) is also reflected in the so-called Causal Markov condition (**CM**). This is a generalization of familiar ideas about screening off, first formulated by Reichenbach (1956), and has figured heavily in recent work on causal inference by Judea Pearl (2000a) and by Peter Spirtes, Clark Glymour, and Richard Scheines ([1993] 2000). **CM** says that, conditional on its parents or direct causes, every variable is independent of every other variable except its effects:

**(CM)** For all  $Y$  distinct from  $X$ , if  $X$  does not cause  $Y$ , then  $Pr(X/Parents(X)) = Pr(X/Parents(X) \cdot Y)$

(for discussion, see Spirtes et al. [1993] 2000; Hausman and Woodward 1999, forthcoming a). There are circumstances under which **CM** fails to hold (see Spirtes et al. [1993] 2000; Hausman and Woodward 1999), but it is a plausible conjecture that in these circumstances, no general test for causation in terms of conditional independence relationships will work. As Hausman and Woodward (1999) argue, insofar as there is any systematic connection between causation and conditional independence relations in acyclic graphs, it appears to be captured by **CM**. **CM** also seems to be equally plausible whether “cause” on its l.h.s. is interpreted as “total cause” or as “contributing cause.” Notice that **CM** is formulated in terms of the parents or direct causes of  $X$ ; this is one reason for thinking that insofar as there is a systematic connection between causation and probabilistic (in)dependence, it will need to appeal to some notion of direct causation.

Contrapositing **CM** gives a sufficient condition for causation in terms of conditional dependence: if  $Pr(X/Parents(X)) \neq Pr(X/Parents(X) \cdot Y)$  for  $Y$  distinct from  $X$  and  $Parents(X)$ , then  $X$  causes  $Y$ . However, this condition is *not* necessary for contributing causation; when there are failures of faithfulness, as in figure 2.3.1 with  $a = -bc$ ,  $X$  will be a contributing cause of  $Y$  even though  $X$  and  $Y$  are uncorrelated, so that  $Pr(X/Parents(X)) = Pr(X/Parents(X) \cdot Y)$ . (One can

think of faithfulness as the converse of the Causal Markov condition; faithfulness says that given a graph and associated probability distribution, the only independence relations are those that follow from the Causal Markov condition alone and not from special parameter values, as in figure 2.3.1 with  $a = -bc$ .) However, provided that the variation in the value of  $X$  is such that if values of  $X$  that are not independent of the value of  $Y$  exist, they are realized, the following condition does seem to be a nontrivial necessary condition for  $X$  to be a contributing cause of  $Y$ <sup>30</sup>: If  $X$  is a contributing cause of  $Y$ , then there is a directed path  $P$  from  $X$  to  $Y$  such that conditional on all the parents of  $Y$  that are not on  $P$ ,  $X$  and  $Y$  are dependent. Note that this condition again requires the notion of a direct cause and hence some framework (graphical or equational) for expressing the way causal factors are connected with each other.

Suppose that we instead take **CC** as a test for whether  $X$  is a total cause of  $Y$ , that is, we take “causes” on the l.h.s. of **CC** to mean “is a total cause.” When understood in this way, the idea underlying requirement (iv)—that one should not control for variables that are causally intermediate between  $X$  and  $Y$ —is correct. Nonetheless, we still need the notion of direct causation to capture what needs to be controlled for. In particular, what needs to be controlled for in testing whether  $X$  is a total cause of  $Y$  is *both* (i) other *direct* causes of  $Y$  that are not on any causal route from  $X$  to  $Y$  and (ii) the *direct* causes of any intermediate variables along any route from  $X$  to  $Y$ . Controlling just for other nonintermediate *total* causes of  $Y$  (besides  $X$ ) is not adequate.<sup>31</sup> To see that (i) is required, consider again the structure in figure 2.3.1 with exact cancellation along the two routes ( $a = -bc$ ). In this structure, to assess whether  $Z$  is a total cause of  $Y$ , we must control for  $X$ . However,  $X$  is not a total cause of  $Y$ . It is  $X$ 's status as a direct cause of  $Y$  that is not intermediate between  $Z$  and  $Y$  that explains why it needs to be controlled for.<sup>32</sup> Thus, the general claim made above about contributing causes is also true of total causes: in characterizing what needs to be controlled for in determining whether  $X$  causes  $Y$ , in either the total or contributing sense, it is not enough to have information about contributing or total causal relationships involving  $X$ ,  $Y$ , and other variables. One also needs information about the direct causal relationships involving those variables.

## 2.5 Causal Claims as Telling Us What Happens under Some (Not All) Interventions

All of the conditions connecting causation and manipulation formulated above (**M**, **TC**, as well as **NC\***, **SC**, and **DC**) make reference to there being “some” possible intervention on  $X$  that, in appropriate circumstances, changes  $Y$ . A natural question is why we should formulate the connection between causation and manipulation this way. Why not replace the “some” with “all,” demanding, for example, that a necessary and sufficient condition for  $X$  to be a total cause of  $Y$  is that, in some set of background circumstances, *all* possible interventions on  $X$  also change  $Y$  or that a necessary and sufficient condition for  $X$  to be a contributing cause of  $Y$  is that (again for some set of background circumstances)

all possible interventions on  $X$  change  $Y$  when off-path variables are held fixed at some (alternatively all) value(s)? There are several reasons this proposed modification is misguided.

First, even if  $X$  is a (total or contributing) cause of  $Y$ , there may be some possible manipulations/interventions on  $X$  (i.e., changes in the value of  $X$  that in other respects meet the conditions for an intervention) that do not change  $X$  “sufficiently” to change  $Y$ . If we represent the causal relationship between  $X$  and  $Y$  as a function  $F$  from  $X$  to  $Y$ , these will be cases in which  $F$  maps two or more different values of  $X$  into the same value of  $Y$ ; that is, cases in which  $F(x_1) = F(x_2) = y_1$  and  $x_1 \neq x_2$ , but in which  $x_3 \neq x_1$  and  $x_3 \neq x_2$  and  $F(x_3) \neq F(x_2)$ . In such cases, a manipulation that changes the value of  $X$  from  $x_1$  to  $x_2$  will not change the value of  $Y$  even though  $X$  causes  $Y$  in the sense that there is some other manipulation of the value of  $X$  ( $= x_3$ ) that will change the value of  $Y$ . Consider the following illustration (2.5.1), which is drawn from Woodward and Hitchcock (2003). A lightbulb is connected to a dial switch, which may be moved through an angle of 0 to 180 degrees. If the switch is moved to any position less than 90 degrees, the light remains off. If the switch is at any position equal to or greater than 90 degrees, this causes the light to come on. On and off are the only two states of the light. There is a causal relationship between the position of the switch and whether the light is on or off, but it is not true that all possible manipulations of the position of the switch will change the state of the light. Moving the switch from 0 to 45 degrees or from 90 to 135 degrees will produce no change in the state of the light. What is true is that there is some possible manipulation of the position of the switch (namely, from less than 90 degrees to greater than or equal to 90 degrees and vice versa) that will change the state of the light. Both **TC** and **M** correctly say that this is sufficient for the relationship to be causal.

This example brings out a more general point that is quite important: the bare claim that  $X$  causes  $Y$  is not very informative. From the perspective of a manipulability account, what one would really like to know is not just whether there is some manipulation of (intervention on)  $X$  that will change  $Y$ ; that is, whether it is true that  $X$  causes  $Y$ . One would also like to have more detailed information about just which interventions on  $X$  will change  $Y$  (and in what circumstances) and exactly how they will change  $Y$ . We may view this sort of detailed information about what will happen to  $Y$  under various hypothetical manipulations of  $X$  as the natural way of spelling out or capturing the detailed content of specific causal claims regarding  $X$  and  $Y$  within a manipulability framework.

There are a variety of devices for conveying such information. One possibility, which we have already made extensive use of, is to write down an explicit mathematical or functional relationship. Consider the relationship (2.5.2)  $F = -k_s X$ , where  $X$  is the extension of a particular type of spring and  $F$  the restoring force it exerts. This does not just claim that extending the spring causes it to exert a restoring force. It also explicitly spells out, in quantitative terms, exactly how various interventions that change the extension in various ways will change the force the spring exerts. Obviously, this more detailed quantitative information is much more useful to an agent who is interested in

manipulation than the nonspecific information that  $X$  causally influences  $F$ . Similarly for other quantitative laws: we may think of the gravitational inverse square law as codifying information about exactly how manipulations that change the distance between two masses or the magnitudes of the masses themselves will change the gravitational force they exert on each other.

Another device that is widely used to convey more specific information about patterns of dependence is contrastive focus. Consider the following variant on an example originally due to J. L. Mackie (1974). A hammer repeatedly strikes chestnuts, always with the same momentum  $m_k$ , as they move along a conveyer belt and causes them to shatter. There is a range of possible momenta  $m_i$  between, say,  $m_1$  and  $m_k$  (i.e.,  $m_1 < m_i < m_k$ ), such that if the hammer were to strike the chestnuts with a momentum in this range, they would still shatter, and similarly for momenta greater than  $m_k$ . For some other range of momenta significantly less than  $m_1$ , striking the chestnuts will not be followed by shattering. Thus, if  $S$  is a variable that can take the possible values  $\{\textit{shattered}, \textit{unshattered}\}$ , then only some and not all possible manipulations of the momentum of the hammer will change the value of  $S$ . As Mackie observes, even though the causal claim that

(2.5.3) The blows cause the shatterings

is true, it would not be true to say, without qualification, that if the actual blows (with momentum  $m_k$ ) had not occurred, the shatterings would not have occurred. One way that the actual blows might not have occurred is that blows might still have occurred but with some momentum that is substantially different from  $m_k$  but greater than  $m_1$ . Given such blows, the shattering still would have occurred.<sup>33</sup>

Some writers (e.g., Bennett 1987) take examples like this to be counterexamples to counterfactual theories of causation. It seems to me that they rather illustrate the desirability of having a device that conveys more specific information about exactly which possible manipulations of a putative cause are being claimed to change an effect (or exactly which alterations of the effect are counterfactually dependent on exactly which alterations of the cause). This is just what being explicit about the contrastive focus of the claim (2.5.3) does. Special circumstances aside, the natural or default interpretation of (2.5.3) is something like this:

(2.5.4) The contrast between (a) the striking of the chestnuts as they actually occurred and (b) situations in which no striking at all occurs causes the contrast between (c) outcomes in which the chestnuts shatter and (d) those in which they do not shatter at all.

Speaking somewhat more compactly and employing the “rather than” locution, we can express this by saying that the (actual) striking of the nuts rather than not striking them at all causes them to shatter rather than not to shatter at all. The device of contrastive focus thus allows us to distinguish between (2.5.4) and

(2.5.5) The (a) actual strikings rather than (e) strikings of slightly greater or slightly less momentum cause the nuts to shatter rather than not to shatter.

This last claim (2.5.5) is not true, although it might be true that

(2.5.6) The occurrence of (a) rather than (e) causes the shatterings to occur in the particular way they do rather than (f) in some slightly different way.

As the reader should be able to see, the use of contrast in these examples maps onto a manipulationist understanding of causal claims in a very direct way. Thus, (2.5.4) conveys the information that a manipulation that changes (a) to (b) will change (c) to (d) and vice versa; (2.5.5) is false because it is not true that manipulations that change (a) to (e) will change (c) to (d); and so on.

I turn now to a second reason for formulating the connection between causation and manipulation in terms of claims about what will happen under “some” rather than “all” interventions on the cause (or at least for using the “all” interventions formulation with a restricted domain of quantification). Return to the example of Hooke’s law (2.5.2)  $F = -k_s X$ . Even if (2.5.2) correctly describes the relationship between some manipulations that change the extension of a particular spring  $s$  (or kind of spring  $S$ ) and the force it exerts, this generalization will almost certainly have “exceptions.” These will be of two different kinds. First, (2.5.2) will likely not hold for *all* manipulations that alter  $X$ . For one thing, if a manipulation extends the spring  $s$  to too great a length, it will no longer exert a linear restoring force. Even within the range of extensions for which (2.5.2) sometimes holds, it will break down for certain ways of extending the spring (e.g., those that involve twisting or distorting it) or under certain background conditions (intense heat). Second, there will be many other springs, different from  $s$  or not of kind  $S$  for which (2.5.2) will not hold. Most of those who use Hooke’s law are aware that it holds only for some and not all manipulations of the spring  $s$ , and also does not characterize the behavior of many springs different from  $s$ , even if they are unable to fully enumerate the conditions under which (2.5.2) fails to hold. Thus, while (2.5.2) describes a genuine causal relationship between extension and restoring force, it would be a mistake to interpret (2.5.2) as a claim about what will happen to  $F$  under “all possible” manipulations of  $X$ , either for all springs or for spring  $s$ . Instead, we should interpret (2.5.2) as making a claim about what will happen under some (range of) manipulations of  $X$ , for some but not all springs, just as (TC) and (M) suggest.

A similar point holds for garden-variety nonquantitative causal claims. Suppose that we are interested in determining whether (2.5.7) ingestion of a dose of some drug  $D$  causes recovery  $R$  from a certain disease in humans. We conduct a suitably designed experiment several times, with a treatment group who suffer from the disease and are given  $D$ , and a control group, also suffering from the disease, from whom  $D$  is withheld. In each experiment we find that everyone in the treatment group recovers and that no one in the control

group does. This is strong evidence that under the conditions that characterize this particular kind of experiment and for these subjects, manipulating whether or not a subject receives *D* changes whether or not the subject recovers, and hence that *D* causes *R*. However, the experimental results do not show and it may not be true that "all possible" manipulations of *D* change whether or not a subject recovers. It may be that the subjects in the experiment would not recover if given *D* under sufficiently different background circumstances and it may also be that there are other subjects who differ sufficiently from those in the experiment who would not recover if given *D*.

This point is recognized in the distinction that is made in the literature on experimental design between *internal* and *external* validity (see, e.g., Cook and Campbell 1979, pp. 50ff). As I understand these notions, the internal validity of an experiment has to do with whether the claim that some factor *C* causes some effect *E* in the background conditions characterizing that very experiment is true. An experiment designed to determine whether *C* causes *E* and in which *C* and *E* are correlated will not be internally valid if, for example, this correlation occurs as a matter of chance or if under these experimental conditions some other factor besides *C* is the only factor that causes *E*. By contrast, external validity has to do, roughly, with the extent to which the causal relationship between *C* and *E* will continue to hold in other circumstances besides those obtaining in the original experiment—with the extent to which the *C-E* casual relationship generalizes to other subjects and background circumstances. If the original experiment was done with nondrinking American males between twenty and thirty, questions of external validity will have to do with, for example, whether the drug will also cause recovery among women, among children or the elderly, among drinkers, and so on. The distinction between internal and external validity is obviously premised on the assumption that it is possible for manipulation of *C* to change *E* and hence for *C* to cause *E* within a certain experimental context and yet for it not to be true that manipulation of *C* changes *E* outside this context. Again, this point supports the claim that **(TC)** and **(M)** ought to be formulated in terms of what will happen under "some" rather than "all" interventions.

The notion of *invariance*, which will receive more extensive treatment in subsequent chapters, provides a convenient way of expressing these points. **(TC)** and **(M)** imply that if a causal relationship between *C* and *E* holds at all, then it must be true that (and the relationship must correctly describe how) for some interventions and background circumstances, *E* will change under those interventions on *C*. This in turn implies that there must be some relationship between *C* and *E* and some interventions on *C* such that if these were to be carried out, that relationship between *C* and *E* would not break down but rather would continue to hold. When this is true, I say that the relationship is invariant under such interventions and background circumstances. Thus, according to a manipulationist account of causation, if a relationship is to qualify as causal, it must be invariant under some interventions. However, as the examples described above show, it is perfectly possible for a relationship to qualify as causal even if it is not invariant under all interventions. The notion

of invariance should thus be understood as relative to a set of interventions and background circumstances: a relationship may be invariant under some interventions and background circumstances but not invariant under others. We can spell out the content of a causal claim more precisely by describing the range or domain of interventions and background circumstances over which it is invariant.<sup>34</sup>

## 2.6 Causation, Counterfactuals, and Reproducibility

Both **(TC)** and **(M)** connect type-causal claims and counterfactuals. **(NC\*)**, one of the conjuncts that makes up **(M)**, says that if  $X$  is a contributing cause of  $Y$ , then a certain counterfactual will be true: there will be some intervention on  $X$  such that, if appropriate other variables were held fixed,  $Y$  (or the probability of  $Y$ ) would change. **(SC)**, the other conjunct making up **(M)**, says that if an existential claim about what would happen under a counterfactual supposition is true—if there is a possible intervention on  $X$  that would change  $Y$  (or the probability of  $Y$ )—then  $X$  causes  $Y$ . **(TC)** and **(M)** thus embody a counterfactual theory of causation, not in the sense that they claim to offer a reductive<sup>35</sup> analysis of causal claims in terms of some (noncausal) notion of counterfactual dependence, but in the sense that they claim that there is systematic connection between causal claims and certain counterfactuals. How should these counterfactuals be understood? A detailed answer to this question must await discussion of the notion of an intervention in chapter 3, but there is a preliminary issue that needs to be addressed. On one way of interpreting the clause “there is some intervention on  $X$  such that if it were to occur, then  $Y$  (or the probability of  $Y$ ) would change,” a sufficient condition for its truth is that there be some single token intervention that changes the value  $X$  and under which the value of  $Y$  (or the probability of  $Y$ ) also changes. Nothing at all is implied about how  $Y$  will respond to interventions on  $X$  on other occasions, no matter how similar to the first.

I specified in section 2.2 that the scope of my discussion was restricted to causal claims that assert the existence of reproducible casual relationships. If there are true type-causal claims, which have implications only for the results of a single manipulation in the manner described in the previous paragraph, they are excluded from my discussion by this restriction. (Independently of this, it is hard to think of realistic examples of such claims and, if there are such examples, they appear to play little role in science.) In accordance with this restriction, I assume that even if a type-causal claim such as (2.5.7) administration of drug  $D$  causes recovery  $R$  is established by single experiment, it should be understood to be generalizable or repeatable in some way; that is, as having some implications about what would happen to the effect variable if the cause variable were appropriately manipulated on other occasions. Thus, the counterfactual claims in **(TC)** and **(M)** about what would happen to  $Y$  were “some” possible intervention on  $X$  to be carried out should be interpreted to mean that there is some intervention on  $X$  such that *if it were*

*possible to intervene to manipulate X repeatedly in that way*, Y (or the probability of Y) would change in some reproducible or repeatable way. (As will become clear in chapter 3, for this counterfactual to be true, I do *not* require that it be technologically or even nomologically or physically possible to manipulate X, either repeatedly or even once.) In other words, the claims about the behavior of Y under some possible intervention on X in **(TC)** and **(M)** are to be understood as implicitly general, in the sense that they are claims about a general or reproducible reaction of Y (or the probability of Y) under interventions on X if it *were* possible to repeat such interventions in appropriate circumstances on other occasions, but not as committing us to the claim that such interventions are physically or practically possible.<sup>36</sup>

As I remarked in section 2.2, the intended sense of “reproducible” is relatively undemanding: the idea is that the response of Y (or the probability of Y) to manipulation of X should be general or stable enough that it is appropriate to speak of manipulating X as a means or strategy for producing Y in some (perhaps highly restricted) circumstances. I take at least this much reproducibility to be implied by the demand that valid experiments must be reproducible. As the discussion in the previous section suggests, reproducibility in this sense carries with it no specific implications about external validity: the response of Y to changes in X might be highly reproducible in a specific experimental context, but not hold at all in different contexts or circumstances. Reproducibility also does not require determinism: if the causal relationship between X and Y is indeterministic in the sense that interventions that change the value of X in the same way do not always produce the same change in the value of Y, although they always produce the same change in the probability distribution of Y, this will count as a reproducible relationship. A similar conclusion will follow if the same manipulation of X merely increases the probability or incidence of Y across some range of circumstances but not others and not always by the same amount.

What, then, does reproducibility rule out? Suppose that (2.6.1) you have a coin that is fair: when tossed, the probability of heads is one half and there are no available variations in the manner of tossing that will change this probability. Only two actions are possible: you may intervene to toss the coin either with your left hand or with your right hand. The coin may land either head or tails. You toss once with your right hand and the result is heads (*H*). The requirement of reproducibility says that this is *not* sufficient to establish the truth of the relevant counterfactuals in **TC** and **M**. That is, given the requirement of reproducibility, we cannot argue that because you intervened to toss the coin with your right (rather than your left) hand and the coin came up heads (rather than tails), it follows that there is an intervention such that if it were carried out, it would change whether (or change the probability of whether) the coin comes up heads.<sup>37</sup> Given the details of the case, there is no reproducible relationship between which hand you use and the results of the toss: you cannot use your choice of hand as a means or strategy for altering the outcome of the tosses. More generally, the requirement of reproducibility implies that it is not a sufficient condition for the truth of claims involving



counterfactuals such as “there is some intervention on  $X$  that would change the value of (or the probability distribution of)  $Y$ ” that there be some single token intervention that changes the value  $X$  and under which the value of  $Y$  (or the probability of  $Y$ ) also changes.

This way of understanding type-causal claims excludes the possibility that there are true type-causal claims that are singularist in content in the sense that it is built into their meaning that they have implications only for what happens in a single instance or a one-shot causal interaction. Of course, a claim like (2.5.7) might be true even if there is only one case (or, for that matter, no cases) in the entire history of the universe in which a realization of  $D$  causes a realization of  $R$ , but if so, this will be a consequence of the contingent fact that only one (or no) realizations of  $D$  happen to occur in the right circumstances for causing  $R$ . What is excluded by the reproducibility requirement is that it is somehow built into our conception of the way the causal relationship between  $D$  and  $R$  works (and is not just a contingent fact about the number of occasions on which  $D$  happens to be realized) that a realization of  $D$  causes  $R$  just once but could<sup>38</sup> never do so again, no matter how closely and often we replicate the circumstances in which that single causing occurs. Moreover, because (as I shall argue in section 2.7) we must appeal to claims about type-causal relationships of the sort embodied in **(M)** to elucidate token-causal claims, the latter are also implicitly general in the same way: contrary to the views of writers like Anscombe ([1971] 1993) and Ducasse (1926), token or singular causal claims always should be understood as committing us to the truth of some type-level causal generalizations.

If we understand the counterfactuals in **TC** and **M** in the generalizable way just described, then one important kind of evidence that is relevant to their truth has to do with whether certain correlations hold when the conditions specified in their antecedents are realized. For example, if the counterfactual claim that  $Y$  would change from a value of  $y_1$  to a value of  $y_2$  if  $X$  were appropriately manipulated from a value of  $x_1$  to a value of  $x_2$  is true for all of the individuals in which we are interested, and if we can observe the  $X$  and  $Y$  values of all those individuals, then we expect that if we were to repeatedly intervene, sometimes setting the value of  $X$  to  $x_1$  and sometimes to  $x_2$ , there will be a (perfect) correlation between these values of  $X$  and the values of  $Y$ . If, instead, the claim that interventions on  $X$  will change the probability distribution of  $Y$  is true (where this probability is strictly between 0 and 1), then there will be a nonperfect correlation between  $X$  and  $Y$ . For example, if it is true that the value of  $R$  would change depending on whether subjects are treated with drug  $D$  or not, then, in an appropriate population, some of whom are so treated and some of whom are not, we should expect to see evidence for the truth of this counterfactual in a correlation between  $D$  and  $R$ : recovery should be more common among those who are treated. Parallel remarks hold for counterfactuals with more complex antecedents, such as the counterfactual in **(M)**. Thus, one way (although certainly not the only way) of testing the counterfactual claims in **(TC)** and **(M)** or of obtaining empirical evidence relevant to their truth will be to carry out the interventions described in their

antecedents and then check to see whether certain correlations hold. This helps to tie down the empirical content of these counterfactuals.

I emphasize that, in this view of the matter, the existence of the correlations just described does not entail and is not entailed by the counterfactuals in **(M)**, but rather, is *evidence* for their truth. Suppose that the coin in (2.6.1) above is flipped ten times with the right hand and, improbably, the result is ten heads. If we count this as a correlation, then it will provide (misleading) evidence for the counterfactual (2.6.2) “If you were to toss this coin with your right hand, the result would be heads,” but that counterfactual (as I interpret it) will nonetheless be false, because the relationship between tossing and outcome is not in the relevant sense reproducible.

An alternative strategy, popular with many philosophers, is to distinguish between sample and population correlations. Roughly speaking, a sample correlation is an observed correlation between  $X$  and  $Y$  in a finite sample passing some standard test for statistical significance; a population correlation is the correlation in the hypothetical population of values of  $X$  and  $Y$  that would result if the experiment of intervening on  $X$  and observing the change in  $Y$  were repeated indefinitely. According to this strategy, the correlation in (2.6.1) is only a sample and not a population correlation, and in the relevant population heads and righthandedness will be uncorrelated. If we follow this strategy, we can say that the truth of the counterfactual (2.6.2) *implies* that an associated population correlation holds and not merely that the correlation is evidence for the counterfactual. In other words, in contrast to the position defended in the previous paragraph, the connection between (population) correlations and counterfactuals is taken to be analytic rather than merely evidential. However, this connection is achieved only by building the modal and hypothetical commitments of the counterfactual (2.6.2) into the notion of a population correlation. As a result, the problems that arise in inferring from correlations to counterfactuals are not solved but simply relocated: they are now problems of inferring from sample to population correlations. An extra term, “population correlation,” has been introduced to mediate between sample correlations and counterfactual claims, but no real reduction of the counterfactual claims in **TC** and **M** to something more epistemically accessible has been achieved. For this reason, though I have no real objection to the population correlation notion, I prefer the more straightforward position described in the previous paragraph.

Thus, in response to the question, Isn’t it possible for  $X$  and  $Y$  to be correlated under interventions on  $X$  but for this correlation to be purely accidental in the sense that it does not result from a causal connection between  $X$  and  $Y$ , but rather occurs because  $Y$  just “happens” to change in value spontaneously in a way that is correlated with the values of  $X$ , my response is that this is indeed possible, assuming that we mean (as I would prefer) by “correlation” something tied to the results of applying statistical tests to observed correlations in finite samples. However, this is not a problem for **TC** and **M** because, as explained above, I interpret the counterfactuals in them as claims about what would happen to  $Y$  under indefinitely many repetitions of

interventions on  $X$ . In the case under discussion in which  $X$  just happens to be correlated with  $Y$ , these counterfactuals will be false, although we will have (misleading) evidence for their truth in the form of a correlation between  $X$  and  $Y$ .

I conclude this section by briefly noting some other respects in which, in real-life experiments, the connection between causal and counterfactual claims, on the one hand, and observed correlations on the other will be far more complex and subtle than in the idealized picture described above. For one thing, in my remarks above, I assumed that *all* the units in the population that was experimentally manipulated shared the same causal structure. In real-life experiments, the manipulated population may be a mixture of different causal structures with different probability distributions over variables of interest. It is well-known that such mixtures may give rise both to “spurious” correlations that do not reflect facts about causal structure and to the nonoccurrence of correlations one might expect to be revealed in facts about causal structure.<sup>39</sup> Moreover, in assuming above that the values of  $X$  and  $Y$  are *observed* or *measured* for all units that are experimentally manipulated, I have idealized away from issues having to do with selection bias and missing measurements. Suppose that in the drug experiment described above, both the value of  $T$  (which measures whether or not a person is treated with drug  $D$ ) and the value of  $R$  (which measures recovery) influence whether or not that person appears in the experimental sample. For example, it may be both that the drug has side effects (e.g., nausea) that influence whether subjects return for measurement of  $R$  and also that those subjects who are feeling particularly sick or well do not bother to return for measurement of  $R$ . In such cases, the observed correlation between measured  $T$  and measured  $R$  can be quite misleading about the actual-causal relationship between  $T$  and  $R$ . Finding procedures for dealing with such bias is a very important practical issue in experimental design. I have ignored such issues not because I regard them as insignificant, but because I wish to focus instead on the connection between experimental manipulation and the meaning or content of causal claims.<sup>40</sup>

## 2.7 Actual Causation

Previous sections have provided an account of so-called type-causal claims of the form  $X$  causes  $Y$  where  $X$  and  $Y$  are variables. What is the relationship between such claims and claims involving actual-, singular-, or token-causal claims, which we may understand as claims to the effect that  $X$ 's assuming some actual value on some particular occasion (for some particular individual) caused  $Y$  to assume some actual value on that occasion? There is no consensus on this issue. As noted above, the philosophy of science and structural equations literature has tended to focus almost exclusively on type-level causal claims and has had relatively little to say about token causation. By contrast, philosophers working in the tradition inaugurated by David Lewis have focused exclusively on token causation, which they have attempted to

treat completely independently of type causation. Moreover, the connection between these two sorts of claims is not transparent. It can be true that smoking causes lung cancer, that Jones smokes, and that Jones develops lung cancer, and yet false that Jones's smoking caused his lung cancer. (It was instead caused by his exposure to asbestos.) Nonetheless, and despite some suggestions to the contrary (e.g., Sober 1985; Eells 1991), it also seems implausible that these two causal notions are unrelated to each other.

In this section, I sketch an account of how the ideas about causal relationships between variables developed above can be extended to capture important features of token causation.<sup>41</sup> My discussion is meant to be illustrative and suggestive, rather than comprehensive and definitive. My intent is to show that the interventionist framework can be used to capture judgments about token causation in an interesting range of cases, but I make no claim to have provided an exhaustive treatment of all possible cases. From my point of view, the general idea of the approach I advocate, which is that we can capture the content of many token-causal claims by means of counterfactuals having to do with what happens under combinations of interventions, is more important than whether the details are correct in every respect. As long as some account along the general lines that I favor can be made to work, a broadly manipulationist approach to token causation will be vindicated.

I work in this section within a quite restricted framework, in which information about deterministic type-causal relationships is assumed to be part of our background knowledge, and the only question is what these type-causal relationships and other background information imply about token-causal relationships. Obviously, there are many cases in which we have knowledge of token-causal relationships even in the absence of knowledge of relevant deterministic type-causal generalizations. For example, I may know with confidence that a blow on the head caused Jones's death, even though I do not know any relevant nontrivial deterministic generalization about the circumstances under which blows on the head are followed by death. Cases of this sort are discussed in chapters 4 and 5.

I begin with an intuitive, but overly restrictive statement of the account that I advocate. This version reproduces some but not all of our judgments about token-causal relationships; for example, it fails to handle cases of symmetrical overdetermination. I then generalize the account to handle such cases. As I have argued above, when we ask whether *X* is a contributing type-cause of *Y*, we are interested in the following question, which I label (T) for future reference:

(T) Is there a directed path between *X* and *Y* such that an intervention on *X* will change *Y* when other variables that are not on this path are held fixed at some possible value?

When we ask this question we are not interested in what the actual values of *X* and *Y* and other variables in the system of interest are, or in whether, given the actual values assumed by other variables that are to be held fixed, an intervention that changes the value of *X* would change the value of *Y*.

Instead, it is enough that there is some intervention that changes some value of  $X$  (not necessarily a change to or from the actual value of  $X$ ) and that will change the value of  $Y$ , given that appropriate other variables are held fixed at some (not necessarily actual) values. In this sense, questions about type causation abstract away from information about the actual values of variables in any particular actual situation and about what would happen under interventions in such situations, given those actual values. By contrast, token causation has to do precisely with such matters. Within the structural equations framework adopted above, we may treat token-causal claims as claims to the effect that some variable  $X$ 's assuming its actual value caused another variable  $Y$  to assume its actual value. We may think of such claims as (or analogues to) type-causal causal claims that are specialized to actual situations in the sense that we are now interested in what will happen to the effect variable when the cause variable changes from its *actual* value to some alternative while various other (appropriate) variables are fixed at some value. The first possibility we will consider is that these other variables are to be fixed at their *actual* values. In particular, rather than considering question (T), we will focus instead on (roughly) the following question (A):

(A) Is there a directed path from  $X$  to  $Y$  such that some intervention that changes  $X$  from its actual value would change  $Y$  from its actual value, when (a) other variables along all other directed paths from  $X$  to  $Y$  are fixed by interventions at their actual values and (b) other direct causes of  $Y$  that are not on any directed path from  $X$  to  $Y$  remain at their actual values?

As an intuitive illustration, assume again that short circuits ( $S$ ), oxygen ( $O$ ), and fire ( $F$ ) are related by equation (2.2.3)  $F = S \cdot O$ , so that there are two causal routes or arrows into  $F$ : one from  $S$  to  $F$  and one  $O$  to  $F$ , as in figure 2.2.2. As argued above, short circuits (type) cause fires (in this case, in both the contributing and total sense) because there is some value of the  $O$  variable ( $O = 1$  or oxygen is present) such that, given that value, intervening to change whether or not a short circuit occurs changes whether or not a fire occurs. Similarly, oxygen (type) causes fires because, when a short circuit is present, intervening to change whether oxygen is present will change whether the fire occurs. Thus, for it to be true that short circuits cause fires, it does not matter whether the actual situation is one in which oxygen is present. By contrast, when we ask, with reference to some concrete situation, whether the short circuit was an actual cause of the fire, we are interested in what the actual values of  $S$ ,  $O$ , and  $F$  are in that situation and whether interventions that change  $S$  from its *actual* value will (or would) change  $F$  from its *actual* value, given the *actual* values of variables on other causal routes, in this case, the actual value of  $O$ . Thus, in a situation in which oxygen is present and a short circuit and fire occur, whether the short circuit is an actual cause of fire will depend on whether, fixing the  $O$  variable at its actual value (oxygen present), an intervention that changes the value of  $S$  from its actual value (in

this case, from present to absent) will change whether the fire occurs. If the causal structure of the system under investigation is correctly represented by (2.2.3), then changing  $S$  under these conditions will indeed change  $O$ , and so the short circuit qualifies as the actual cause of the fire. By contrast, in a situation in which oxygen is absent, (2.2.3) tells us that the fire will not occur (indeed, it will fail to occur regardless of the value of  $S$ ), so that in this circumstance, the short circuit cannot be the cause of either the occurrence or nonoccurrence of the fire. (As we will see, to assume that the causal structure of the situation is correctly represented by (2.2.3) is to assume that various possible preempting causes of the fire are not present. If they are present, (2.2.3) does not correctly represent the situation we are trying to model and we require some more complicated equation or set of equations.)

Pursuing this basic line of thought, we can construct a “first pass” account of token causation by devising an analogue to **(DC)** and **(M)** in which references to possible values of variables are replaced by references to actual values. This suggests the following procedure for determining whether  $X$ 's taking of some value ( $X = x$ ) is an actual cause of  $Y$ 's taking some actual value ( $Y = y$ ). First, draw the causal graph connecting  $X$  to  $Y$ , guided by the characterization of direct causation **(DC)**. Then determine whether the following two conditions are satisfied:

**(AC)**

(AC1) The actual value of  $X = x$  and the actual value of  $Y = y$ .

(AC2) There is at least one route  $R$  from  $X$  to  $Y$  for which an intervention on  $X$  will change the value of  $Y$ , given that other direct causes  $Z_i$  of  $Y$  that are not on this route have been fixed at their actual values. (It is assumed that all direct causes of  $Y$  that are not on any route from  $X$  to  $Y$  remain at their actual values under the intervention on  $X$ .)

Then  $X = x$  is an actual cause of  $Y = y$  if and only if both conditions **(AC1)** and **(AC2)** are satisfied.

To illustrate **(AC)**, consider first the well-known desert traveler case.  $A$  poisons the water in the traveler's ( $T$ ) canteen with cyanide.  $B$  punctures a hole in the canteen and, as it happens, all of the water drains out before  $T$  has a chance to drink and he dies of dehydration.  $B$ 's action was the actual cause of death, even though it is true that if  $B$  had not punctured the canteen,  $T$  would have died anyway from drinking the poisoned water. Let  $P$  be a variable that takes the value true or false depending on whether  $A$  poisons  $T$ 's canteen,  $H$  be a variable that takes the values true or false depending on whether  $B$  punctures a hole in  $T$ 's canteen,  $D$  a variable taking the values true or false depending on whether  $T$  is dehydrated,  $C$  a variable taking the values true or false depending on whether  $T$  ingests cyanide, and  $M$  a variable taking the values true or false depending on whether  $T$  dies. The causal diagram for this situation is shown in figure 2.7.1.<sup>42</sup>

The corresponding equations are:

$$(2.7.1) \quad C = P \cdot \neg H$$

$$(2.7.2) \quad D = H$$

$$(2.7.3) \quad M = C \vee D$$

The actual values of the variables are  $H = \text{true}$ ,  $P = \text{true}$ ,  $C = \text{false}$ ,  $D = \text{true}$ , and  $M = \text{true}$ . The justification for the diagram and equations is as follows. There are values of the other variables (any set of values in which  $H = \text{false}$ ) such that if they are held fixed at those values and the value of  $P$  is changed by an intervention, the value of  $C$  will vary. Hence,  $P$  is a direct cause of  $C$ . Similarly,  $H$  is a direct cause of  $C$  because, if  $P$  is fixed at true, changing the value of  $H$  by an intervention will change the value of  $C$ . In particular, for cyanide to be ingested it must be the case both that  $P$  is true and that  $H$  is false, which is what (2.7.1) says. Similarly,  $H$  is a direct cause of  $D$  because, regardless of the values of the other variables, whether or not  $D$  is true depends on whether or not  $H$  is true. There are no other direct causes of  $D$ , because, given the causal setup, nothing else changes the value of  $D$  for either value of  $H$  (or for any of the values of the other variables). Thus,  $D$  depends only on  $H$ , as (2.7.2) says. Finally, it is obvious that both  $C$  and  $D$  are direct causes of  $M$  and that if either is true,  $M$  will be true, as (2.7.3) says.

Let us now assess whether  $H$ 's taking the value true was an actual cause of  $M$ 's taking the value true.  $H$  and  $M$  are both true, and hence condition AC1 is satisfied. In applying condition AC2, consider the route from  $H$  to  $M$  that goes through  $D$ . There is just one direct cause of  $M$  that lies off this route, namely,  $C$ . Fixing  $C$  at its actual value ( $C = \text{false}$ ), changing the value of  $H$  from true to false will change the value of  $M$  and hence AC2 is satisfied.  $H$ 's being true is thus an actual cause of  $M$ 's being true.

By contrast,  $P$ 's being true is *not* an actual cause of  $M$ 's being true because condition AC2 is not met. There is a single route from  $P$  to  $M$  and one direct

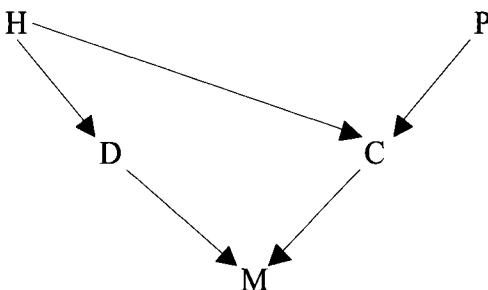


Figure 2.7.1

cause of  $M$  that is not on this route, namely,  $D$ . When we hold  $D$  fixed at its actual value,  $D = \text{true}$ , changing the value of  $P$  does not change the value of  $M$ .

We can put all of this somewhat more simply in terms of claims about what will (or would have) happen(ed) under possible interventions as follows. Given that, in the actual situation,  $T$  did not ingest cyanide, an intervention that changes whether or not a hole was punctured would have changed whether or not he died. By contrast, given that  $T$  was dehydrated, changing whether or not cyanide was placed in his canteen would not change whether he dies. This asymmetry of counterfactual dependence captures or cashes out the content of our saying that the puncturing and not the poisoning was the actual cause of  $T$ 's death.

Finally, note that this example provides an additional illustration of the observation made at the beginning of this section: it is possible for  $X$  (putting poison in a canteen) to be a type cause of  $Y$  (death), for instances  $x$  of  $X$  and  $y$  of  $Y$  to occur, and yet for  $x$  to fail to be an actual cause of  $y$ . However, this does not show that actual causation and type causation are unrelated notions. Instead, type-causal relationships play a fundamental role in elucidating actual-causal relationships, and both notions embody a counterfactual or manipulationist notion of causation.

Next, consider an example from Hall (forthcoming). A boulder falls, causing a hiker to duck. If he had not ducked, he would not have survived. Our intuitive judgments are that although the fall causes him to duck and the ducking causes survival, the fall does not cause survival. The example thus illustrates an apparent failure of transitivity involving actual causation. Letting  $F$  represent whether or not the boulder falls,  $D$  whether or not the hiker ducks, and  $S$  whether or not he survives, we have the causal diagram in figure 2.7.2 and the associated equations:

$$(2.7.4) \quad D = F$$

$$(2.7.5) \quad S = -F \vee D$$

with the actual values of the variables being  $F = \text{true}$ ,  $D = \text{true}$ ,  $S = \text{true}$ .

On the account of actual cause given above, ducking is clearly an actual cause of survival: both  $D$  and  $S$  are true, and hence condition (AC1) is met. Moreover, fixing  $F$  at its actual value, an intervention that changes the value of  $D$  will make a difference to the value of  $S$ , so that condition (AC2) is met as well. Similarly, the falling of the boulder is an actual cause of ducking.

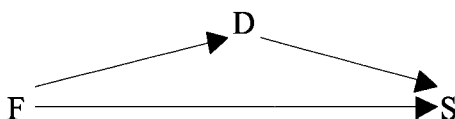


Figure 2.7.2



Is the falling an actual cause of survival? Although condition (AC1) is met, condition (AC2) is not met. If we consider the direct route from  $F$  to  $S$  and fix the intermediate variable ( $D$ ) along the other route at its actual value, changing  $F$  will not change  $S$ . The only other route is the direct route from  $F$  to  $S$ , and because there are no intermediate variables along this route, there is no well-defined operation of fixing such variables in order to evaluate the influence of  $F$  on  $S$  along the  $F$ - $D$ - $S$  route. Thus, in agreement with intuition, the falling of the boulder is not an actual cause of the hiker's survival. (This result depends, of course, on the assumption that there are no intermediate variables on the route from  $F$  to  $S$  that does not go through  $D$ .)

This example provides an additional illustration of the difference between the claim that the value of  $X$  was an actual cause of the value of  $Y$  and the claim that  $X$  (type) causes  $Y$ . Although the falling boulder is not an actual cause of the hiker's survival, it is nonetheless true that  $F$  type causes  $S$  (in the contributing cause sense) in structures of this sort. We assume this when we draw the causal diagram or write down the equations (2.7.4)–(2.7.5), but it also follows from the definition (DC) of direct cause.  $F$  is a direct cause of  $S$  because when the only other variable  $D$  in the system assumes the value ( $D = \text{false}$ ), changing  $F$  makes a difference to  $S$ . In other words,  $F$  qualifies as a direct and a contributing (type) cause of  $S$  because it is true that if the hiker does not duck, whether or not the boulder falls makes a difference to whether he survives. However, the fall of the boulder does not qualify as an actual cause of survival, because given the actual value of  $D$  (hiker ducks), the hiker would survive whether or not the boulder falls. By contrast, in a world in which the boulder falls and the hiker fails to duck and dies, the fall of the boulder would be an actual cause of his failure to survive.

As I noted above, this treatment of the boulder example depends crucially on the absence of any intermediate variable on the direct route from  $F$  to  $S$ . This raises the obvious question of why it wouldn't be equally or more correct to include such a variable in our representation of the example. Consider, for example, the variable  $B$  representing the presence or absence of a boulder at an intermediate point along the (spatial) route from the point of fall to the hiker's body but sufficiently close to the hiker that he lacks the time to duck if he has not done so already. If we interpolate such a variable along the (previously) direct route from  $F$  to  $S$  and fix it at its actual value ( $B = \text{present}$ ), then (it would seem) an intervention that changes the value of  $F$  will change the value of  $S$  (along the route that goes through  $D$ ), and hence the fall of the boulder will qualify as an actual cause of survival. This in turn illustrates a more general point about the use of equations and graphs to which I have already alluded: conclusions about causal relationships are sensitive to one's choice of representation.

Although the matter deserves more detailed treatment than I can give it here, I believe that if the situation being modeled has the structure originally described, there are good reasons for *not* including  $B$  in the representation. Let me first make explicit a point that was only hinted at in the previous paragraph: if the interpolation of  $B$  is to make any difference to the conclusions we

reach, *B* must not itself affect *D*: if there is an arrow from *B* to *D*, the analysis given above applies all over again, with *B* replacing *F*, and we are led to the conclusion that neither *B* nor *F* is an actual cause of *S*. If *B* does not affect *D*, this must be because *B* describes the position of the boulder at some late point in its trajectory (say, two meters from the hiker's body) such that if the hiker has not already ducked at this point, there is too little time for the boulder's being in that position to cause ducking before the boulder strikes. Thus, in evaluating whether the boulder's falling caused survival, we must freeze *B* at its actual value (boulder present two meters from hiker) and then ask whether if *F* were to change from its actual value (i.e., if the boulder were not to fall), *S* would change. However, this is, at the very least, a puzzling counterfactual, given the original description of the situation. How could the boulder both not fall and be two meters from the hiker's head? If we are to sensibly entertain this counterfactual, we must imagine a situation with a different causal structure. Consider, for example, a situation in which there are two different ways a boulder might be present at *B*, either by falling (in which case, falling influences *D* because the hiker sees the boulder coming) or in some other way that does not influence *D*. To make this second possibility concrete, imagine that an enemy throws boulders at the hiker from behind his back. They don't cause the hiker to duck because he doesn't see them coming. Suppose now that enemy throws a boulder at the hiker while at the same time another boulder falls from the original source, causing the hiker to duck. Because of this, both boulders miss and the hiker survives. If the second boulder had not fallen from the original source, the hiker would not have ducked, would have been hit by the enemy's boulder, and would not have survived. Under this scenario, the counterfactual "If a boulder were two meters from the hiker's head and if the (second) boulder had not fallen, then the hiker would not have survived" makes sense and is true. By the same token, it now seems correct to say that the fall of the boulder caused the hiker to survive. However, this is a situation with a very different causal structure from the original situation envisioned above. Given the structure of the original situation, there is no way that a boulder can strike the hiker other than by falling.<sup>43</sup>

As a final illustration of **AC**, consider a case of trumping (Schaffer 2000b; Lewis 2000). The sergeant and the major give orders to the corporal. The major's orders always trump the sergeant's, in the sense that the corporal always does what the major orders, regardless of the sergeant's orders. But when the major gives no orders, the corporal always follows the sergeant's orders. Suppose that the major and sergeant order "Advance" and the corporal advances. A common judgment is that the major's ordering "Advance" causes the corporal to advance and that the sergeant's orders do not cause the corporal to advance.<sup>44</sup> Suppose that we accept this judgment, at least provisionally. Several writers (see, e.g., Schaffer 2000b) claim that this is a case of causation without counterfactual dependence, on the grounds that although the major's order causes the advance, if the major had given no order, the corporal still would have advanced because he would have followed the sergeant's orders.

Even if we agree that the major's order but not the sergeant's causes the corporal to advance, this last claim (that the example involves causation without counterfactual dependence) is not true (at least for the notion of counterfactual dependence defended here) and provides a striking illustration of how thinking about causation within a framework of (binary) "events" that either occur or fail to occur often obscures rather than illuminates what is really going on. It is important to our judgment about the above example that there are more than just two possibilities for what the major, sergeant, and corporal may do, or that if there are only two possibilities, neither of them is "give no orders," for otherwise, the asymmetry between the major and the sergeant will disappear.<sup>45</sup> To take the simplest alternative, let us suppose that there are three possibilities available to the major: he may either order "Advance" ( $M = 1$ ), give no order at all ( $M = 0$ ), or order "Retreat" ( $M = -1$ ). Similarly for the sergeant's orders and for the corporal, who may either advance ( $= 1$ ), do nothing ( $= 0$ ), or retreat ( $= -1$ ). Writing  $S$  for the sergeant's orders and  $C$  for the corporal's response, we may represent the situation by means of the following equations:

$$(2.7.6) \quad \text{If } M \neq 0, C = M$$

$$(2.7.7) \quad \text{If } M = 0, C = S$$

Fixing  $S$  at its actual value  $S = 1$ , we see that according to **AC**,  $M = 1$  is an actual cause of  $C = 1$ , because an intervention that changes the value of  $M$  to  $-1$  would change the value of  $C$  to  $-1$ .  $S = 1$  is not an actual cause of  $C = 1$  because fixing  $M$  at its actual value of 1, there are no interventions on the value of  $S$  that would change the value of  $C$ .

Although **AC** seems to deliver the intuitively correct verdict in the cases just described, it fails to do so in other cases. One of the simplest involves symmetric causal overdetermination. Suppose that each of two events  $c_1$  and  $c_2$  are individually causally sufficient for  $e$  (i.e., each causally determines that  $e$  will occur). For example, each of two campers throws a lighted cigarette into the forest, where each cigarette on its own would have produced a forest fire, and a fire follows, or each of two marksmen shoots a prisoner who then dies, with each shot being sufficient on its own for death. Although some philosophers claim otherwise, it seems to me that our intuitive judgment in such cases is that both events,  $c_1$  and  $c_2$ , are causes of  $e$ . If we let  $A = 1$  or 0 according to whether  $c_1$  occurs,  $B = 1$  or 0 according to whether  $c_2$  occurs, and  $C = 1$  or 0 according to whether  $e$  occurs, then the equation corresponding to the above examples is simply

$$(2.7.8) \quad C = \max(A, B)$$

Fixing  $A$  at its actual value  $= 1$  in accord with **(AC)**, we see that changing the value of  $B$  from its actual value ( $B = 1$ ) does not change the value of  $C$ . So, according to **AC**,  $c_2$  ( $B = 1$ ) is not an actual cause of  $e$  ( $C = 1$ ). By parity of reasoning,  $c_1$  is also not a cause of  $e$ .

If we accept that this is the wrong answer, how might we change or extend **AC** to cover such cases? A very natural thought is that the causal dependence of  $e$  on  $c_1$  fails to express itself in counterfactual dependence only because  $c_2$  happens to be present as well and that at least part of the grounds for judging that  $c_1$  causes  $e$  is that we can see that if we were somehow to remove the influence of  $c_2$  on  $e$ , then whether  $e$  occurs would indeed depend on whether  $c_1$  occurs. This suggests relaxing (AC2) to allow the other direct causes  $Z_i$  of  $Y$  that are not on the route from  $X$  to  $Y$  and the intermediate variables that are on other routes from  $X$  to  $Y$  to be fixed at values other than their actual value.  $X = x$  will then count as an actual cause of  $Y = y$  if and only if (AC1) is satisfied and there is a route from  $X$  to  $Y$  such that by freezing the direct causes of  $Y$  that are not on this route at some possible (not necessarily actual) value, an intervention on  $X$  will change the value of  $Y$ . Thus, in the example represented by (2.7.8),  $A = 1$  is a cause of  $C = 1$  because there is a value of  $B$ —namely,  $B = 0$ —for which fixing  $B$  at that value and intervening to change the value of  $A$  will change the value of  $C$ .

A moment's thought, however, shows that this proposed modification is too permissive in the sense that it ends up counting noncauses as causes. Return to equation (2.2.3)  $F = S \cdot O$  in which  $S$  is a variable representing the occurrence/nonoccurrence of a short circuit,  $O$  a variable representing the presence or absence of oxygen, and  $F$  a variable representing the occurrence or nonoccurrence of a fire. Suppose that in the actual situation oxygen is absent, the short circuit present, and the fire does not occur. There is a possible value of  $O$ , namely,  $O = present$ , for which an intervention that changes whether the short circuit occurs will change whether the fire occurs. However, contrary to what the proposed modification of **AC** suggests, the presence of the short circuit is not an actual cause of the absence of the fire.

The solution to this difficulty that seems best on balance is due to Halpern and Pearl (2000) and Hitchcock (2001a). Consider a particular directed path  $P$  from  $X$  to  $Y$  and those variables  $V_1 \dots V_n$  that are not on  $P$ . Consider next a set of values  $v_1 \dots v_n$ , one for each of the variables  $V_i$ . The values  $v_1 \dots v_n$  are in what Hitchcock calls the *redundancy range* for the variables  $V_i$  with respect to the path  $P$  if, given the actual value of  $X$ , there is no intervention that in setting the values of  $V_i$  to  $v_1 \dots v_n$ , will change the (actual) value of  $Y$ . The actual values of the variables  $V_i$  are, of course, in the redundancy range with respect to  $P$  but nonactual values of the variables  $V_i$  will also be in the redundancy range if, given the actual value of  $X$ , we can set the variables  $V_i$  to those values without disturbing the actual value of  $Y$ . In the symmetric overdetermination example represented by (2.7.8),  $B = 0$  is in the redundancy range of  $B$  with respect to the directed path from  $A$  to  $C$ , because given the actual value of  $A$ ,  $A = 1$ , the value of  $C$ ,  $C = 1$  would be unchanged if  $B = 0$ . By contrast, in the short circuit example (2.2.3) with  $S = present$ ,  $O = present$ , and  $F = present$ , the value  $O = absent$  is not in the redundancy range with respect to the path from  $S$  to  $F$ , because given the actual value of  $S$ , the fire will not occur if  $O = absent$ . The Halpern–Pearl–Hitchcock proposal is basically to modify AC2 in **AC** to allow fixing the other direct causes  $Z_i$  of  $Y$  that are not on a given

route  $P$  at possible, nonactual values only when those values of  $Z_i$  are within the redundancy range of  $Z_i$  with respect to  $P$ . Thus, what I propose for dealing with cases of symmetric overdetermination is to replace **(AC)** with **(AC\*)**:

(AC\*1) The actual value of  $X = x$  and the actual value of  $Y = y$ .

(AC\*2) For each directed path  $P$  from  $X$  to  $Y$ , fix by interventions all direct causes  $Z_i$  of  $Y$  that do not lie along  $P$  at some combination of values within their redundancy range. Then determine whether, for each path from  $X$  to  $Y$  and for each possible combination of values for the direct causes  $Z_i$  of  $Y$  that are not on this route and that are in the redundancy range of  $Z_i$ , whether there is an intervention on  $X$  that will change the value of  $Y$ . (AC\*2) is satisfied if the answer to this question is “yes” for at least one route and possible combination of values within the redundancy range of the  $Z_i$ .

$X = x$  will be an actual cause of  $Y = y$  if and only if (AC\*1) and (AC\*2) are satisfied.<sup>46</sup>

Thus, in the symmetric overdetermination example (2.7.8), the value  $B = 0$  is, as we have seen, within the redundancy range of  $B$  with respect to the path from  $A$  to  $C$  and, setting  $B = 0$ , an intervention that changes  $A = 1$  to  $A = 0$  will change the value of  $C$  from 1 to 0. Hence, AC\*2 is satisfied and  $A = 1$  is an actual cause of  $C = 1$  by **AC\***.

What is the intuition between allowing off-route direct causes to be fixed at nonactual values as long as these are within the redundancy range of those variables? As we have seen, in cases involving symmetric overdetermination, the relationship between each of the overdetermining cause variables and the effect is such that given the actual value of the cause, the counterfactual dependence of the effect on the other overdetermining cause variable is masked. Fixing one of the cause variables at nonactual values allows the actual-causal relationship involving the other cause to reveal itself in a relationship of counterfactual dependence. For example, we see that both  $A = 1$  and  $B = 1$  are causes of  $C = 1$  in a case of symmetric overdetermination by noting that if  $B = 1$  had not occurred ( $B = 0$ ),  $C$  would be counterfactually dependent on  $A$ , and that similarly,  $C$  would be counterfactually dependent on  $B$  if  $A$  did not occur. However, because our interest is in actual token-causal relationships, we cannot allow just any variable to be set at just any nonactual value. In particular, it seems clear that it would defeat the whole point of the enterprise if we permitted settings of variables to nonactual values that (according to the model we are considering) have the consequence that any actual-causal relationships are arbitrarily disrupted. For example, we want to rule out setting the off-route direct causes of the effect of interest to possible values that are inconsistent with the value actually taken by the effect. AC\*2 permits us to consider nonactual values of some variables but avoids the problem just described by restricting the nonactual values we consider to those within the redundancy range of the variables in question, because by definition such changes will not influence the value of the effect.

My interest in this section has been in showing how the apparatus of directed graphs and a manipulationist approach to causation can be used to reconstruct commonsense judgments about token-causal relationships. I want to conclude, however, on a somewhat more skeptical note. If the discussion in this section has been successful, what it has accomplished is successfully isolate facts about patterns of counterfactual dependence, as revealed in hypothetical manipulations, that are relevant to commonsense token-causal judgments and causal distinctions. However, in at least some of the cases discussed above, it is controversial what the deliverances of common sense are and even more so whether (or even what it would mean to say that) such deliverances are “correct.” The suggestion I want to make is that to the extent that commonsense causal judgments are unclear, equivocal, or disputed, it is better to focus directly on the patterns of counterfactual dependence that lie behind them—the patterns of counterfactual dependence are, as it were, the “objective core” that lies behind our particular causal judgments, and it is such patterns that are the real objects of scientific and practical interest.

To illustrate this idea, consider again the difference between, on the one hand, a case of preemption, in which one assassin shoots ( $c_1$ ) and causes the death of a victim ( $e$ ), but a second, backup assassin would have shot ( $c_2$ ) and killed him if the first had not and, on the other hand, a case of symmetric overdetermination in which both assassins kill the victim simultaneously. According to the account presented in (AC\*), the relevant difference is this: in the preemption case, (1) given the actual state of the alternative cause ( $c_2$  fails to occur), the occurrence of  $e$  is counterfactually dependent on the occurrence of  $c_1$ ; by contrast, in an otherwise similar case of symmetric overdetermination, the relevant pattern of counterfactual dependence is (2) if (contrary to what actually happened)  $c_2$  had failed to occur, then if  $c_1$  had occurred,  $e$  would have occurred, and (still assuming that  $c_2$  does not occur) if  $c_1$  had failed to occur,  $e$  would not have occurred. Thus in these circumstances,  $e$  would be counterfactually dependent on  $c_1$ . Similarly, if  $c_1$  had not occurred,  $e$  would be counterfactually dependent on  $c_2$ . In the discussion above, I assumed that the most common judgment regarding cases of symmetric overdetermination (hence the judgment we want our account to reproduce) is that both  $c_1$  and  $c_2$  cause  $e$ . By contrast, David Lewis (1986b, appendix E) suggests that common sense fails to deliver a clear verdict in such cases and hence that it is no defect in his account that it yields the conclusion that neither  $c_1$  nor  $c_2$  causes  $e$ . My guess is that Lewis is wrong about common sense,<sup>47</sup> but it also seems to me that in an important respect it does not matter much whether we count  $c_1$  and  $c_2$  as causes of  $e$  in this case as long as we can agree about what the relevant patterns of counterfactual dependence are. In particular, suppose that one holds that  $c_1$  is a bona fide cause of  $e$  when the pattern of counterfactual dependence connecting  $e$ ,  $c_1$ , and  $c_2$  is as in (1) above, but that when the pattern is (2), the relationship between  $c_1$  and  $e$  is not causation, but rather something else (call it “causation\*”). The difference between this position and the view described by (AC\*) looks largely verbal, as long as both parties agree that the relevant patterns of counterfactual dependence that distinguish the

two cases are indeed (1) and (2).<sup>48</sup> There is, of course, still the interesting *descriptive* question of how people use the word “cause” in connection with cases of symmetric overdetermination and what accounts will reproduce this usage, but it is not clear that there is any further sense, over and above this, to the question of whether  $c_1$  “really” causes  $e$  in cases of symmetric overdetermination.

Similarly for several other cases discussed above. Some may regard trumping as a case of preemption and others may regard it as a case of overdetermination (cf. n. 44). But if, as I assume is the case, we agree about all of the relevant counterfactuals, it is not clear that there is a further question over and above these about whether the sergeant’s orders “really” cause the corporal to advance or not, unless that question is simply a (misleading) way of asking about the judgments that most people in fact endorse in such a case. This assessment seems particularly compelling to the extent that our concern is with causal explanation, for once we have been given information about the complete patterns of counterfactual dependence in the symmetric overdetermination and trumping cases as well as a description of the actual course of events, it appears that nothing has been left out that is relevant to understanding why matters transpired as they did.

## 2.8 Causation, Omissions, and Serious Possibilities

In our discussion of the falling boulder example in 2.7, we rejected the idea that it was appropriate, given the causal structure of this example, to consider the possibility that the boulder both failed to fall and yet (somehow) appeared a few meters from the hiker’s head. It was not that this was (in itself) a logical or causal or nomological impossibility, but rather that, to take this possibility seriously, we needed to consider an example with a rather different causal structure from the one we originally set out to analyze, one in which some independent mechanism or process, other than falling, is responsible for the appearance of the boulder in close proximity to the hiker. At least in ordinary contexts, the possibility that the boulder both fails to fall *and* appears near the hiker’s head *and* doesn’t get there as a result of following a trajectory from some independent source but instead, say, simply materializes near the hiker’s head is not one that we are prepared to take seriously.

In this section, I want to explore in more detail this idea of a “possibility that we are willing to take seriously” and the way it influences our causal judgments. Let me begin with an example (much discussed recently) due to Michael McDermott (1995, p. 525):

Suppose that I reach out and catch a cricket ball. The next thing along in the ball’s direction of motion was a solid brick wall. Beyond that was a window. Did my action prevent the ball hitting the window? (Did it cause the ball to not hit the window?) Nearly everyone’s initial intuition is, “No, because it wouldn’t have hit the window irrespective of whether you had acted or not.” To this I say, “If the wall had not been there, and I had not acted, the ball would have hit the window. So between us—me

and the wall—we prevented the ball hitting the window. Which one of us prevented the ball hitting the window—me or the wall (or both together)?” And nearly everyone then retracts his initial intuition and says, “Well, it must have been your action that did it—the wall clearly contributed nothing.”

John Collins (2000) compares this example with another of similar structure in which the wall is replaced by a second fielder. The second fielder would have caught the ball if the first fielder had not, but as it happens, the first fielder does catch it. Collins claims, I think correctly, that whether or not we agree with McDermott’s suggestion about the first version of the example, we are considerably *more* willing to accept the judgment that the first fielder prevented the shattering in the second version of the example (in which the second fielder rather than the wall is present) than in the first version. Why is this? After all, in both cases, the interaction of the ball and first fielder and the failure of the ball to reach the window are the same; whether the wall and or the second fielder is present seem equally irrelevant and extrinsic to the interaction that does take place between the ball and first fielder.

Collins traces this difference in our reactions to what we are willing to take as a serious possibility in the two examples. In the second example, it seems not at all far-fetched to imagine that (contrary to what we originally stipulated) if the first fielder had missed the ball, the second would have too; fielders often do miss balls, and we know that whether or not they do so depends in a very sensitive way on many complicated factors having to do with timing, coordination, attention, environmental conditions such as wind, and so on. Small differences in these can lead a fielder to miss a catch. By contrast, as Collins remarks, it seems “more far-fetched” to suppose that the brick wall somehow disappears or that the ball somehow passes right through it (2000, p. 29). At least in the example as originally described, without trick Rube Goldberg devices that might realize these possibilities, they are not seen as “serious.” The point becomes clearer as we imagine additional versions of the example in which the destruction of the window involves even more far-fetched possibilities. Imagine, for example, that the window is made of shatter-proof glass and protected by a barrier consisting of three feet of solid steel, and consider the possibility that during the flight of the ball, very quickly acting workmen might have removed the barrier and treated the glass with a special chemical that would have restored its vulnerability to shattering, although in actual fact no such workmen were present and no one possesses such a chemical (although it exists and “might” have been discovered). Here I take it that all will agree that the first fielder’s catch did not prevent the shattering; regardless of whether the first fielder catches the ball, there is no serious possibility that the shattering would occur. Unlike the two-fielder example, no one is tempted to argue that the first catch did prevent the shattering on the grounds that if the first fielder missed and the barrier had been removed and the glass treated, the window would have shattered.

Examples of this sort are readily multiplied. Suppose that a particular doctor *D* is in sole charge of a hospital patient’s care, and that standard medical



practice, well-known to *D*, mandates administration of a readily available antibiotic at the first sign of fever. Administration of the antibiotic in the early stages of fever will virtually always eliminate it. Through inattention, *D* fails to observe that the patient has developed a fever, omits to administer the antibiotic, and the patient dies. Here it seems very natural to judge that *D*'s failure to administer the antibiotic caused the patient's death. The manipulability account of causation of course supports this judgment: manipulating the values of the variable *A* reflecting whether or not the antibiotic is administered will change the value of the variable *S* that reflects whether or not the patient lives or dies. Now contrast this with a second case. *X* lives at a great distance from the patient, has no responsibility for his care, is not aware that he exists, and is not a doctor. Suppose that if *X* had happened to go the hospital room where the patient is being cared for and had realized that the patient was developing a fever and had learned that administration of the antibiotic was the appropriate response and had administered the drug, the patient would have survived: the patient's survival is counterfactually dependent on whether *X* does these things. Nonetheless, we are not ordinarily inclined to think that *X*'s failure to do these things caused the patient's death.

Again, the difference between these cases seems connected to our willingness to take seriously certain possibilities and not others. That doctor *D* might have administered the antibiotic strikes us as a serious possibility: it is his responsibility to do this, the antibiotic is available, and so on. Most doctors administer the antibiotic on similar occasions, and probably *D* himself has usually or always done so on other occasions. By contrast, although it is logically and, we may suppose, causally or nomologically possible for *X* to administer the antibiotic (no law of nature prohibits *X* from flying to the distant city, wandering into the patient's hospital room, and administering the antibiotic), there seems no reason at all to take this possibility seriously, and as a result we don't judge *X*'s failure to do these things as a cause of the patient's death.

What sorts of considerations influence whether we are willing to take a possibility seriously? Obviously, the extent to which an outcome represents a serious possibility is a matter of degree, to which a number of considerations are relevant. As the above examples illustrate, the probability of an event's occurring, given the actual obtaining background conditions (or perhaps those that usually or commonly obtain in similar situations) is one relevant consideration. However, it is plainly not the only relevant consideration; considerations having to do with moral requirements, expectations, and custom may matter. (The doctor is obligated to treat the patient, and for this reason, his doing so strikes us as a serious possibility even if he is almost always irresponsible.) Similarly, considerations having to do with whether an outcome is controllable at all (or easily or cheaply controllable) by current technology may also matter. It seems unlikely that there is any algorithm for determining whether a possibility will be or ought to be taken seriously.

What implications do these observations have for manipulability accounts of causation? It may seem that they pose a serious problem. The two versions of the example in which the ball moves toward the window seem to share the

same counterfactual structure. In both versions, an event  $c_1$  occurs (the first fielder catches the ball), a second event  $c_2$  (the collision with the wall, the second fielder catching the ball) does not occur, and a third event  $e$  (the shattering of the window) also does not occur. In both cases, (i) if  $c_1$  had not occurred,  $c_2$  would have occurred and  $e$  would not have occurred, and (ii) if  $c_1$  had not occurred and  $c_2$  had not occurred,  $e$  would have occurred. However, we are more willing to regard  $c_1$  as a cause of  $e$  in the second scenario (in which the second fielder rather than the wall is present) than in the first scenario. Thus, it may seem that our causal judgments are not fixed just by our beliefs about which counterfactuals (even interventionist counterfactuals) are true. Moreover, as the above examples and many others illustrate, it appears that if we don't appeal to some notion like that of serious possibility, a manipulability theory will at the very least be led to causal judgments that are different from those that are ordinarily accepted. For example,  $X$ 's failure to administer the antibiotic will count as a cause of the patient's death, the failure of a large meteor to strike me as I write these words will count as a cause of my writing them, and so on. On the other hand, it may seem that the notion of a serious possibility is a rather unclear and subjective notion, and that introducing it into one's theory of causation will make that theory similarly unclear and subjective.

Let me begin with the worry about subjectivity. As I have already intimated, I think that it is true that in some cases an investigator's (or investigative community's) interests and purposes (and not just how the world is) influence the possibilities that are taken seriously; for example, in biomedical and engineering contexts, whether or not some possibility is, as a practical matter, controllable by available technology may influence whether it is taken seriously (see chapter 5).<sup>49</sup> On the other hand, as the examples described above illustrate, at least some of the considerations that go into such decisions are based on facts about how the world operates that seem perfectly objective. For example, there is nothing arbitrary or subjective about the claim that boulders don't materialize out of thin air or that balls do not pass through solid brick walls. Similarly, it is a straightforward matter of fact that strangers from distant cities rarely wander into hospital rooms and administer life-saving medication.

A second point is that if the reliance of a manipulability theory on judgments about which possibilities are to be taken seriously is a flaw, it is a flaw that is shared by many other theories of causation. In most theories of causation, the properties, variables, equations, and state spaces that the theorist uses to model or represent specific cases will either directly reflect judgments about which possibilities are to be taken seriously or will end up doing largely the same work as such judgments.<sup>50</sup> For example, there is certainly a regular association between the failure of strangers to come to the aid of patients in circumstances like those described above and the patients' deaths. Regularity theorists who hold that  $X$ 's failure does not cause the patient's death will need to find some way of ruling out such regularities as a basis for causal attribution. Attempts to do this by arguing, for example, that certain omissions or

absences are not bona fide events or do not involve the exemplification of genuine properties will end up tracking at least some features associated with the notion of serious possibility. In counterfactual theories such as Lewis ([1973] 1986b, 2000), judgments that are related to judgments about serious possibility will surface in judgments about the extent of “influence” (see chapter 5 for additional discussion) and arguably in judgments about closeness or similarity of possible worlds. Sensitivity of causal judgments to choice of representation will also surface in judgments that certain kinds of events (e.g., overly disjunctive events<sup>51</sup>) do not exist and hence cannot figure in causal relationships (as in Lewis 1986d, p. 190) and in judgments about the fragility or individuation of events (cf. 193ff).

What about the point concerning similarity of counterfactual structure but divergence of causal judgment? In my view, the most natural way of thinking about the above examples is this: the structures of counterfactual dependence that we find in them hold independently of which possibilities we are willing to take seriously; in this sense, they are interest-independent and objective. For example, whether or not we are willing to take *X*'s intervention as a serious possibility, it is nonetheless true that if *X* had intervened, then the patient's life would have been saved. Similarly, the counterfactual claims that, in the first version of the window example, the window would have shattered if the fielder had missed and the wall was absent, and that the window would not have shattered if the fielder had missed and the wall was present are true or false, independently of what anyone is willing to take as a serious possibility. By contrast, causal judgments reflect both objective patterns of counterfactual dependence and which possibilities are taken seriously; they convey or summarize information about patterns of counterfactual dependence among those possibilities we are willing to take seriously. In other words, to the extent that subjectivity or interest relativity enters into causal judgments, it enters because it influences our judgments about which possibilities are to be taken seriously. However, once the set of serious possibilities is fixed, there is no further element of arbitrariness or subjectivity in our causal judgments; relative to a set of serious possibilities or alternatives, which causal claims are true or false is determined by objective patterns of counterfactual dependence. Thus, relativizing causal judgments to a set of serious possibilities (or, what I take to be the same thing, to the choice of some system of representation that reflects those possibilities) does not introduce subjectivity everywhere or indiscriminately but rather, at most, introduces it in a constrained or limited way.

In this view of the matter, although counterfactuals concerning nonserious possibilities can be straightforwardly true or false, to the extent that the possibilities that figure in them are nonserious, they do not guide our causal judgments. Instead, our causal judgments will be influenced just by those counterfactuals we think are true and that concern only serious possibilities. Thus, to the extent that we think that *X*'s intervention to save the patient's life is not a serious possibility, we do not allow the true counterfactual claim that if the intervention had occurred, the patient's life would have been saved to yield the judgment that *X*'s failure to intervene caused the patient's death.

Similarly, to the extent that we are unwilling to regard it as a serious possibility that the ball somehow would have got by the wall and reached the window if the first fielder missed it, the counterfactuals (i) and (ii) described above that involve these possibilities are not reflected in (do not serve as a basis for) our causal judgment. The counterfactual that does serve as a basis for our causal judgment concerning the wall example is instead one that involves serious possibilities only: whether or not the fielder caught the ball, the window would not have shattered (because the wall is present); hence, the catch did not prevent the shattering. When the second fielder rather than the wall is present, our view of what is a serious possibility shifts, and this accounts for the divergence in causal judgment between the two examples.<sup>52</sup>

I conclude with a remark about the general attitude we should take toward causal claims that involve omissions or preventings, as in the second of the examples above. Some writers take the view that these can never be genuine or *bona fide* or literal causes (and perhaps that they cannot be effects as well). Some theories of causation (such as theories that see causation as a matter of the transfer of energy and momentum) may lead to this conclusion. My contrary view, for which I argue in chapter 5, is that we need to recognize that at least some omissions and preventions are genuine causes and can figure in causal explanations—not just in ordinary life, but in science as well. In my view, it is a virtue, not a defect, in the manipulability theory that it allows for this possibility, and a defect in transfer of energy and momentum theories if they do not. If we also wish to say that some variables, values of which correspond to omissions (such as the failure of *X* to help the patient), are not causes even though changes in those values will change the values of other variables, we will need some independent grounds for excluding these. The “not a serious possibility” rationale just described provides one such ground.

## 2.9 Causation as a Cluster Concept

I turn now to another objection to the manipulability account. According to this objection, “cause” is a “cluster concept” involving a number of different criteria for its application. From this perspective, a manipulability theory arbitrarily singles out just one of these criteria—whether the relationship between *C* and *E* is potentially exploitable for purposes of manipulation—at the expense of all of the others, each of which is equally relevant to whether a relationship should be labeled as causal. A particularly clear statement of this point of view is provided by Brian Skyrms (1984). Skyrms suggests that the causal claims made in ordinary, macroscopic contexts typically satisfy many or all of the criteria associated with the various theories of causality proposed by philosophers; for example, it is typically true that all and only causes are statistically relevant for their effects when one controls for relevant background conditions, that effects are counterfactually dependent on their causes, that if one “wiggles” a cause variable experimentally the effect variable will wiggle in response, that causes are linked by law to their effects, that cau-

sation involves the transfer of energy and momentum from cause to effect, and that causes are linked to their effects by spatiotemporally continuous processes. According to Skyrms, because these different criteria tend to be satisfied together in ordinary contexts, we developed a concept of causation that is (or was) an “amiably confused jumble” of or a “cluster concept” involving all of these criteria. This jumble is “amiable” because, in most situations, we don’t have to confront the issue of what to do when some of the criteria are satisfied but not others. However, recent scientific developments have changed this. In particular, in the relationship between the spin states of the separated particle pairs in the EPR experiment, we have a situation in which some of the criteria are satisfied but not others. On the one hand, arguing in support of the claim that the spin state  $S_1$  of one of the separated particle pairs causes the spin state  $S_2$  of the other, we have the fact that  $S_1$  raises the probability of  $S_2$ , that on a very natural interpretation of “counterfactual dependence,”  $S_2$  is counterfactually dependent on  $S_1$ , and that  $S_1$  and  $S_2$  are linked by a generalization that looks like a “law.” On the other hand, one cannot, by manipulating or intervening on  $S_1$ , alter or change  $S_2$  (or vice versa), and there is no spatiotemporally continuous process linking  $S_1$  to  $S_2$ . Skyrms concludes that in this case, “the old cluster concept loses its heuristic value and becomes positively misleading” (1984, p. 254). He does not explicitly say that there is no truth of the matter about whether  $S_1$  “really” causes  $S_2$ , but this would be a natural conclusion to draw from his remarks: the relationship between  $S_1$  and  $S_2$  satisfies some but not all of the commonly accepted criteria for causation, and whether we conclude that  $S_1$  causes  $S_2$  will depend on which criteria we decide to weight most heavily. This conclusion is explicitly drawn by Richard Healey (1994).

What can be said in defense of the way a manipulability theory privileges one of the criteria for causation over the others? First, unlike Skyrms, I think that the criteria do not just come apart in *recherché* cases such as EPR. Instead, they conflict in many more ordinary, macroscopic situations, and when they do, it is always or virtually always the manipulability criterion that wins out. To the extent that this is so, it is a mistake to think of causation as a cluster concept in which all of the criteria described above have approximately equal weight. For example, there are many perfectly ordinary cases, involving causation by omission (discussed in section 2.8) or by double prevention (5.11), in which there is no spatiotemporally continuous process linking a cause to its effect and no transference of energy and momentum, but which are correctly counted as cases of causation by the manipulability theory. Similarly, the claim that if  $C$  and  $E$  are statistically dependent when one controls for relevant background conditions, then  $C$  causes  $E$  gives mistaken results in a number of situations, including all of the cases described by Hausman and Woodward (1999) in which the Causal Markov condition fails. Consider a well-known example, due to Salmon (1984), in which a cue ball simultaneously strikes the 3 and 8 balls, sending each into opposite pockets. It may well be true that the sinking of the 8 ball is statistically relevant to the sinking of the 3 ball, even after one has controlled for the other causally relevant factor (the collision with

the cue ball), but the sinking of the 8 ball does not cause the sinking of the 3 ball. Here again, the manipulability condition captures this causal judgment very nicely: we readily agree that interfering with the path of the 8 ball after the collision so that it does not go into the pocket will not alter whether the 3 ball goes into the pocket, and it is on this basis that we conclude that the sinking of the 8 ball does not cause the sinking of the 3 ball. A parallel point holds for counterfactual dependence as a criterion for causality: as argued above, counterfactual dependence fails as a criterion for causation in many ordinary cases unless the counterfactuals in question are interpreted along interventionist lines.

A second consideration is this: whatever the appeal of the cluster concept account as a description of the concept of causation with which we ordinarily operate, it is a problematic concept from the point of view of methodology—it is not a concept we *should* adopt. On the one hand, if we formulate the cluster theory in such a way that satisfaction of all of the above criteria is necessary for the application of the concept “causation,” we will exclude a large number of scientifically interesting cases of causation. On the other hand, if we say that “most” or “many” of the criteria must be satisfied or that some criteria are more “important” than others or that “different criteria will be weighted differently in different contexts,” then unless we can explain with some precision what the quoted phrases mean, we will end up with a concept of causation that is vague and unclear, and the application of which to specific cases is uncertain and contestable. We will also have a notion that lacks a clear motivation or point; that is, we will have no answer to the question of why the particular criteria that make up our notion of causation are (or should be) grouped together, or why we should not add various criteria to this list or remove others or assign different weights to the criteria we use. (Again, the response that certain criteria capture “our” concept of causation simply invites the question of why those criteria should not be replaced with others, which characterize a different, perhaps more useful concept, e.g., causation\*, which dispenses with spatiotemporal continuity but retains some of the other criteria described above.) In this connection, it is worth reminding ourselves that it is often far from obvious, particularly in the biomedical, behavioral, and social sciences and in history, what researchers mean when they claim that various relationships are causal. Even if some version of the cluster concept idea captures the imprecise and jumbled way that some researchers use the word “cause” in these disciplines, some additional argument is required to explain why we should want such a concept. One of many virtues of a monocriterial view like the manipulability theory is that it forces investigators to be less vague and noncommittal about what they mean when they use this word.

## Interventions, Agency, and Counterfactuals

My discussion in chapter 2 relied on an intuitive understanding of the notion of an intervention. In this chapter, I first provide a more detailed characterization of this notion and then explore a number of additional issues raised by the manipulability approach to causation. I also consider its relationship to traditional agency theories of causation and to David Lewis's counterfactual theory.

### 3.1 Interventions Characterized

As remarked in chapter 2, it is heuristically useful to think of an intervention as an idealized experimental manipulation carried out on some variable  $X$  for the purpose of ascertaining whether changes in  $X$  are causally related to changes in some other variable  $Y$ . However, although the notion of an idealized experiment suggests an activity carried out by human beings, we will see shortly that any process, whether or not it involves human activities, will qualify as an intervention as long as it has the right causal characteristics. The idea we want to capture is roughly this: an intervention on some variable  $X$  with respect to some second variable  $Y$  is a causal process that changes the value of  $X$  in an appropriately exogenous way, so that if a change in the value of  $Y$  occurs, it occurs only in virtue of the change in the value of  $X$  and not through some other causal route.

#### 3.1.1 *An Example*

To focus our intuitions, return to the experiment designed to determine whether treatment with a drug causes recovery from a disease. Assume that we have a population of subjects, all of whom have the disease. Each subject is assigned a treatment that consists in the subject's either receiving or not receiving the drug. We may thus represent the treatment received by an individual subject  $u_i$  by means of a binary variable  $T$  that takes one of two values 1 and 0, depending, respectively, on whether  $u_i$  does or does not receive the drug. Whether or not a subject receives the drug is entirely determined by the value of the treatment variable he is assigned. Similarly, recovery may be represented

by means of a variable  $R$  taking values 0 and 1, depending on whether or not individuals with the disease recover. It is assumed that there is a sense (to be elucidated below) in which both variables  $T$  and  $R$  are capable of taking the values 0 and 1 for each individual subject. That is, any subject who receives the drug might not have done so and vice versa. Similarly, it is possible for the same subject to either recover or not recover, (perhaps depending, for example, on whether he receives the drug).

Intuitively, what we want to know is whether, if various experimental subjects  $u_i$  who have not received the treatment and who suffer from the disease (for whom  $T(u_i) = 0$  and  $R(u_i) = 0$ ) were to be given the drug (i.e., if  $T(u_i)$  were to be changed to 1), they would recover or would be more likely to recover (whether  $R(u_i)$  would be changed to 1 or whether the probability that  $R(u_i) = 1$  would increase). Obviously, we cannot investigate this question by both giving the treatment to and withholding it from the same subject. Hence, we employ a more indirect method: we divide the subjects with the disease into two groups, one that receives the drug and the other (the control group) that does not, and then observe the incidence of recovery in the two groups. The experimenter's interventions (which we may represent by means of an intervention variable  $I$ ) thus consist in the assignment of values of  $T$  to individual subjects.

### 3.1.2 *Interventions Characterized Informally*

What conditions should the experimenter's interventions meet in the ideal case? In asking this question, my intention is *not* to inquire about the most general set of conditions that are necessary for any experiment to provide information about the efficacy of the drug. Many imperfect or nonideal experiments can provide information about the efficacy of the drug if the right conditions are satisfied, as can nonexperimental investigations. Instead, my interest is in formulating a notion of intervention that fits with the project pursued in chapter 2, that of providing truth conditions for claims about total, direct, contributing, and actual causal relationships (as reflected in **M**, **TC**, and **AC\***) by appealing to facts about what would happen under interventions. I emphasize that this semantic or interpretive project is very different from the project of specifying the full range of conditions under which causal claims can be inferred from statistical information, a project that has been ably pursued by other writers and to which I do not aim to contribute. The remarks that follow about ideal interventions thus should not be read as claiming that *only* when a manipulation of  $X$  with respect to  $Y$  meets my conditions for an intervention can we infer anything about the causal relationship between  $X$  and  $Y$ , but rather as an attempt to find a notion of intervention according to which claims like **M**, **TC**, and **AC\*** are both plausible and not unilluminatingly circular. I will add that if the reader wishes to have a concrete picture in mind of the notion of intervention that I am attempting to capture, the obvious candidate is randomized experiments. Randomized experiments are not the only way of learning about causal relationships, but they are certainly *a* way and they can be very reliable.



Given this project, I suggest that the following conditions are appropriate. First, the value of the treatment variable for each individual subject should be determined entirely by the experimenter's interventions (i.e., by the value of the intervention variable  $I$ ). Suppose, for example, that prior to the setting up of the experiment, whether or not subject  $u_i$  takes the drug depends on the value of some endogenous variable  $V$  ( $V$  might represent access to medical care or some personality variable, such as willingness to seek medical help). In carrying out the experiment, the introduction of the intervention variable  $I$  must "break" this existing causal connection between  $V$  and  $T$ , so that the value of  $T$  is now set entirely exogenously by  $I$  and is no longer influenced by  $V$ . This means, for example, that subjects are not allowed to determine on their own whether they will take the drug but that this is instead determined entirely by the experimental design adopted by the experimenter.

We may capture this aspect of the experiment by means of the idea that the intervention variable acts as a *switch* for the variable intervened on. Suppose that  $X$  is an endogenous variable to which certain other variables  $V$  (the endogenous causes of  $X$ ) are causally relevant and that  $I$ , our candidate intervention variable, is also causally relevant to  $X$ . Because the variables  $V$  are causally relevant to  $X$ , this means that there are some values of  $I$  for which some changes in the values of the variables in  $V$  will change  $X$ . Nonetheless, there may be other values of  $I$  such that there are no changes in the values of the variables in  $V$  that will change the value of  $X$ . Instead, for these values of  $I$ , the value of  $X$  is a function of the value of  $I$  alone. In this case,  $I$  is a switch for  $X$  with respect to  $V$ . As an (imperfect) illustration, drawn from Woodward and Hitchcock (2003), consider a stereo receiver that has dials for volume, treble, and bass and a power button. The setting of each is causally relevant to the frequency and amplitude of the soundwaves that are emitted from a speaker. However, the settings make a difference to the sound emitted only when the power button is in the "on" position. When the power button is set to "off," there are no possible changes in the position of the other dials that will change the output of the speaker. Under this condition, it is as though the causal connection between the position of the other dials and the output of the speaker is broken. The position of the power button is thus a switch for the output with respect to the other dials.

In this case, there is only one value of the power button that renders the settings of the other dials irrelevant, and this allows us to set the output to just one setting ("off"). If we imagine instead that the stereo has an additional dial, the positions of which allow us to fix the speaker output at any level we wish, regardless of the positions of the other dials, we have something closer to what we would like in an intervention variable: the variable should be such that it allows an experimenter to fix the value of some target variable  $X$  at a variety of different levels in a way that makes it insensitive to changes in the value of other variables that previously causally influenced  $X$ .

More generally, a variable  $S$  acts as a switch for  $X$  with respect to  $V$  if and only if there is some value of  $S$  for which changes in  $V$  will change  $X$  and some

other value of  $S$  (ideally, a range of values for  $S$ ) for which changes in  $V$  will not change  $X$  and for which the value of  $X$  is determined by the value of  $S$  alone. Informally, the value of the switch variable “interacts” with the value of  $X$  in such a way that when the switch is in certain positions, the causal connection between  $X$  and  $V$  is “broken.” We can use this notion to capture the idea that when an intervention on  $X$  occurs, the value of  $X$  (e.g., level of drug in the bloodstream in the experiment described above) is entirely determined by the intervention; the previously existing endogenous causal connections (e.g., voluntary decisions by subjects about whether to take the drug) that have determined the value of  $X$  in the past no longer do so. The notion of an intervention as a switch captures some of the features of the “arrow-breaking” conception of interventions, advocated by writers such as Pearl (2000a) and Spirtes et al. ([1993] 2000) and described in chapter 2, while avoiding some of the aspects of this idea that some philosophers (e.g., Cartwright, 2003) have found objectionable. When the switch is “on” and the value of  $X$  sensitive only to the value of  $I$ , we may think of this as a matter of breaking all other arrows previously directed into  $X$ , replacing them with a single arrow directed from  $I$  to  $X$ .

A second condition the experimental manipulation should meet is that if the experimenter’s interventions  $I$  are correlated with certain other causes of recovery besides  $T$  (whether because  $I$  is caused by these other causes of recovery or for any other reason), this will undermine the reliability of the experiment. This would happen, for example, if the patients in the treatment group had, on average, stronger immune systems than those in the control group. However, it would be too strong to require that  $I$  (or  $T$ ) be uncorrelated with *all* other causes of  $R$ . As long as  $T$  is efficacious,  $I$  and  $T$  will be correlated with any other causes of  $R$  that are themselves caused by  $I$  or by  $T$ . For example, if treatment by the drug does cause recovery and does so by killing ( $K$ ) a certain sort of bacterium, then it will be no threat to the validity of the experiment if the experimenters’ interventions  $I$  are correlated with  $K$ , even though  $K$  causally affects  $R$ . What we need to rule out is the possibility that there are causes of  $R$  that are correlated with  $I$  or caused by  $I$ , and that affect  $R$  independently of the  $I \rightarrow T \rightarrow K \rightarrow R$  causal chain. Moreover, for reasons that should be familiar from section 2.4, the word “causes” in this requirement that “other causes of  $R$  (that are not on the directed path that goes through  $T$ ) cannot be correlated with or caused by  $I$ ” must be interpreted to mean “contributing causes” rather than “total causes.”<sup>1</sup>

A third condition is that  $I$  should not affect recovery independently of  $T$  but only, *if at all*, through it. In other words, *if* there is a directed path from  $I$  to  $R$ , it must go through  $T$ . The italicized phrases in the preceding sentences are included to explicitly allow for the possibility that  $I$  does not cause  $R$  because  $T$  does not cause  $R$ . That is, this third condition does not require that  $T$  cause  $R$  (this would be to introduce a genuinely vicious circularity into the characterization of an intervention; see below), but only that *if*  $I$  causes  $R$ ,  $I$  must do so in virtue of being a cause of  $T$ . Among other things, this third condition implies that  $I$  must not be a common cause of both  $T$  and  $R$ . This condition

would be violated if, for example, administration of the drug was by shotgun blast (assuming that this would directly adversely affect recovery). Less fancifully, it would be violated if the subjects learn whether they have been assigned to the treatment group or the control group and this makes those in the treatment group more hopeful and those in the control group more discouraged and this in turn has an effect on whether they recover, which is independent of any effects of the drug per se. In this case too,  $I$  directly affects  $R$  independently of  $T$ , and we will not be able to reach reliable conclusions about the effect of  $T$  on  $R$ .

### 3.1.3 Interventions Characterized More Precisely:

#### **IV and IN**

In assembling these requirements to characterize the notion of an intervention, it will be useful to proceed in two steps. First, we will characterize what it is for  $I$  to be an *intervention variable* for  $X$  with respect to  $Y$ .<sup>2</sup> This is a type-level causal notion. Second, we will then use this to characterize the notion of an *intervention* on  $X$  with respect to  $Y$ : this is a token-level causal notion.

Let  $X$  and  $Y$  be variables, with the different values of  $X$  and  $Y$  representing different and incompatible properties possessed by the unit  $u$ , the intent being to determine whether some intervention on  $X$  produces changes in  $Y$ . Then  $I$  is an intervention variable for  $X$  with respect to  $Y$  if and only if  $I$  meets the following conditions:

#### **(IV)**

11.  $I$  causes  $X$ .
12.  $I$  acts as a switch for all the other variables that cause  $X$ . That is, certain values of  $I$  are such that when  $I$  attains those values,  $X$  ceases to depend on the values of other variables that cause  $X$  and instead depends only on the value taken by  $I$ .
13. Any directed path from  $I$  to  $Y$  goes through  $X$ . That is,  $I$  does not directly cause  $Y$  and is not a cause of any causes of  $Y$  that are distinct from  $X$  except, of course, for those causes of  $Y$ , if any, that are built into the  $I$ - $X$ - $Y$  connection itself; that is, except for (a) any causes of  $Y$  that are effects of  $X$  (i.e., variables that are causally between  $X$  and  $Y$ ) and (b) any causes of  $Y$  that are between  $I$  and  $X$  and have no effect on  $Y$  independently of  $X$ .
14.  $I$  is (statistically) independent of any variable  $Z$  that causes  $Y$  and that is on a directed path that does not go through  $X$ .

(“Cause” in this characterization always means “contributing cause” rather than “total cause.”)

Given the notion of an intervention variable, an *intervention* may be defined as follows:

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

Before turning to elucidation of these conditions, let me return to an example from chapter 2 that represents a potential complication. Suppose that  $C$  is a common cause of both  $X$  and  $Y$  with no additional causal relationship between  $X$  and  $Y$ . In particular, the causal relationships among these variables conform to the equations

$$X = aC$$

$$Y = bC$$

Suppose that we carry out a manipulation of  $X$  with respect to  $Y$  that alters the value of  $b$ ; that is, it alters the *relationship* between  $C$  and  $Y$ . Under this manipulation, call it  $M$ , the value of  $Y$  will change for fixed values of  $X$ . If we count this as a legitimate intervention on  $X$  with respect to  $Y$ , then applying, for example, **TC**, we will reach the mistaken conclusion that  $X$  causes  $Y$ . Thus, we want to avoid counting  $M$  as an intervention. There are at least two ways of doing this. The first is to count this as a case in which, although  $M$  changes  $X$ , it also changes  $Y$  via a route that does not go through  $X$  in violation of condition I3 above. This is not particularly ad hoc; after all, in changing the relationship between  $C$  and  $Y$ ,  $M$  in effect induces a change in  $Y$  for fixed values of  $C$  and does so in the absence of any causal connection between  $X$  and  $Y$ . A second possible strategy is to add an additional clause I5 to **IV**, which would read as follows:

15.  $I$  does not alter the relationship between  $Y$  and any of its causes  $Z$  that are not on any directed path (should such a path exist) from  $X$  to  $Y$ .

Note that determining whether I5 is satisfied does *not* require that we (already) know whether there is a directed path from  $X$  to  $Y$ ; instead, it requires only that we be able to identify other causes  $Z$  of  $Y$  that are off this path and to determine whether the putative intervention on  $X$  changes the relationship between  $Z$  and  $Y$ . It is realistic to suppose that our background knowledge often enables us to do this; recall the example from chapter 2 in which our background knowledge assures us that setting the position of the barometer dial by consulting a randomizing device does not alter the relationship between atmospheric pressure and storm occurrence.

In what follows, I adopt the first strategy, that of counting examples of the sort under discussion as violations of I3. This is the simpler strategy and allows us to avoid introducing an additional complication into **IV**. Readers who are unhappy with this choice may incorporate I5 into **IV** instead.

### 3.1.4 The Conditions **IN** Illustrated Graphically

We may use directed graphs to provide a more intuitive representation of what the conditions I1–I4 involve. I1 tells us that if  $I$  is an intervention variable for  $X$  with respect to  $Y$ , then there is a directed path from  $I$  into  $X$ ; I2 tells us that for

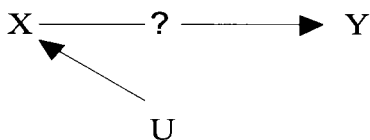


Figure 3.1.1

some values of  $I$ , the value of  $X$  depends only on the value of  $I$ . For those values, the introduction of the intervention variable thus “breaks” any other arrows directed into  $X$ . If we represent the claim that  $X$  causes  $Y$ , the truth or falsity of which we are trying to establish, by means of an arrow from  $X$  to  $Y$  punctuated with a question mark, then when an intervention variable  $I$  for  $X$  takes one of its “on” values, any variable  $U$  (distinct from  $I$ ) that was previously a cause of  $X$  is no longer such a cause. The result of introducing the intervention variable  $I$  is thus to replace the structure in figure 3.1.1 with the structure in figure 3.1.2. This is illustrated by the example considered above, in which the result of the experimenter’s intervention  $I$  is that the treatment variable  $T$  (corresponding to  $X$ ) is no longer caused endogenously by the variable  $V$  (corresponding to  $U$ ) but rather just by  $I$ .

Condition I3 is designed to rule out causal structures like that in figure 3.1.3 or figure 3.1.4. Condition I4 is designed to rule out structures like that in figure 3.1.5, where the line without arrowheads between  $I$  and  $Z$  represents the fact that  $I$  and  $Z$  are correlated, for whatever reason. If we are willing to accept a version of the so-called common cause principle that says that a correlation between two variables can arise only because one causes the other or they have some common cause or causes, then structures like figure 3.1.5 will arise only because one of the structures in figure 3.1.6 holds, where  $W$  represents one or more common causes of  $I$  and  $Z$ .

If we assume the common cause principle, then an intervention variable may be characterized graphically as follows. An intervention variable  $I$  for  $X$  with respect to  $Y$  may be represented by a directed arrow from  $I$  into  $X$  meeting the following three conditions: (i) the arrow from  $I$  to  $X$  is the only arrow directed into  $X$ ; (ii) any directed path from  $I$  to  $Y$  goes through  $X$ ; and (iii) if there is a directed path from any other variable  $V$  into  $I$ , then any directed path from  $V$  to  $Y$  goes through  $X$ .

These ideas about the graphical representation of interventions may be further illustrated by the *ABS* system considered in section 2.2, in which atmospheric pressure  $A$  is a common cause of a barometer reading  $B$  and the



Figure 3.1.2

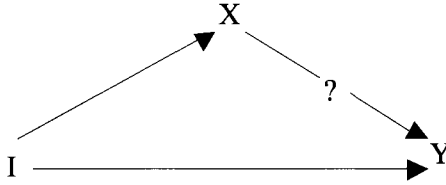


Figure 3.1.3

occurrence of a storm  $S$ . As noted, when we intervene to fix the value of  $B$  according to the results generated by a random-number generator, this breaks the previously existing endogenous causal relationship between  $A$  and  $B$ : the value of  $B$  is now determined by the intervention variable  $I$  rather than by  $A$ , as condition (I2) requires. It is thus natural to think of this as a matter of replacing the structure in figure 3.1.7 with the structure in figure 3.1.8.

This provides an additional illustration of interventions breaking (all) arrows directed into the variable intervened on, replacing them with a single arrow directed from the intervention variable into the variable intervened on.

By contrast, suppose that an intervention occurs on  $A$ . If figure 3.1.7 correctly represents the causal structure of the system, then the relationships between  $A$  and  $B$  and between  $A$  and  $S$  should be preserved under some interventions on  $A$ . In other words, if the causal structure 3.1.7 is correct, the result of some interventions on  $A$  should be representable as in figure 3.1.9. More generally, if a directed graph correctly represents a system of causal relationships, arrows directed *out* of a variable intervened on should be preserved or remain unbroken under at least some interventions. If we have a graph in which  $X$  is represented as a direct cause of  $Y$  and yet every intervention on  $X$  with respect to  $Y$  appears to break the arrow from  $X$  to  $Y$  (in the sense that any correlation between  $X$  and  $Y$  disappears under such interventions), no matter what the values at which other variables are held fixed, then the alleged causal relationship between  $X$  and  $Y$  is not genuine. (This is just **NC\*** restated in graphical terms.)

A similar point holds for other parts of the manipulated system. For example, assuming that the causal link or mechanism connecting  $A$  and  $S$  is genuinely

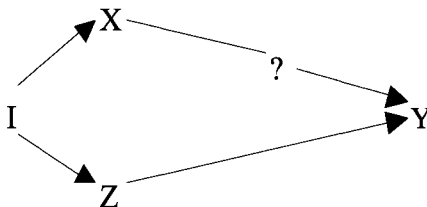


Figure 3.1.4

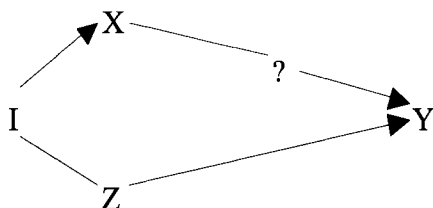


Figure 3.1.5

distinct from the link or mechanism connecting  $A$  and  $B$ , then an intervention on  $B$  will not change either the value of  $A$  or the relationship between  $A$  and  $S$ ; that is, it will not break the arrow from  $A$  to  $S$ . More generally, it is *all and only* arrows directed into the variable intervened on that are broken. All other arrows, both those directed out of the variable intervened on and arrows directed out of and into other variables, are preserved intact. We thus see that built into the idea of an intervention and its graphical representation is the thought that it changes some things and leaves other things unchanged or invariant. In effect, the causal structure in a diagram like figure 3.1.7 spells out in detail what will be changed and what will be left unchanged under various hypothetical interventions. Again, this illustrates the central idea of a manipulability account of causation, which is that we can cash out the content of causal claims in terms of what they imply about what will happen under hypothetical interventions.

As a further illustration of this idea, suppose that the correct causal structure is given by figure 3.1.10 rather than by figure 3.1.7; that is, there is a direct causal link from  $B$  to  $S$ , as well as links from  $A$  to  $B$  and to  $S$ . Then, as before, any intervention on  $B$  should break the arrow from  $A$  into  $B$  (the value of  $B$  is now fixed by  $I$  rather than by  $A$ ). But now, for some such intervention  $I^*$ , the causal connection from  $B$  to  $S$  should be preserved, and as a result, the association between  $B$  and  $S$  should also be preserved rather than disrupted (see figure 3.1.11). In other words, in contrast to figure 3.1.8, there should be some intervention on  $B$  that changes  $S$ . As the reader may check, interventions on all other variables should produce the same results in the case of both figure 3.1.7 and 3.1.10. The difference in causal structure between figure 3.1.7 and figure 3.1.10 can be thus cashed out in terms of a difference in what will happen under some intervention on  $B$ : some such intervention on  $B$  will change  $S$  if figure 3.1.10 is correct, but if figure 3.1.7 is correct, no intervention on  $B$  will change  $S$ .

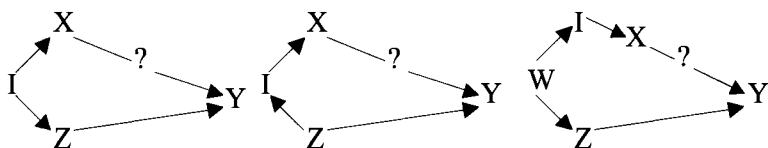


Figure 3.1.6

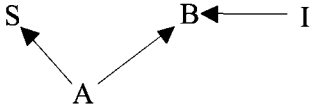


Figure 3.1.7

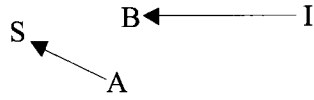


Figure 3.1.8

### 3.1.5 Further Elucidation of *IN*

I turn now to a more detailed discussion of *IN*.

**Relativity** First, note that the notion of an intervention on  $X$  is defined only relative to a second variable  $Y$ . There is no such thing as an intervention on  $X$  simpliciter.  $I$  may thus qualify as an intervention on  $X$  with respect to  $Y$  but not as an intervention on  $X$  with respect to some other variable  $Z$ . This should be unsurprising. If what we are interested in is whether there is a causal relationship between  $X$  and  $Y$ , then an intervention on  $X$  must have the right sort of relationship to  $Y$ ; for example, it must not directly cause  $Y$ . Obviously, an intervention might meet this condition with respect to  $Y$  but not with respect to  $Z$ .

**Nonanthropomorphism** Second, although there will be realistic cases in which manipulations carried out by human beings will qualify as interventions in virtue of satisfying *IN*, the conditions in *IN* make no reference to human activities or to what human beings can or can't do. Notions such as "human agency" and "freely chosen action" do not occur as primitives in *IN*. Instead, the conditions in *IN* are characterized purely in terms of notions such as "cause" and (statistical) "independence." An event or process not involving human action at any point will qualify as an intervention on  $X$  as long as it satisfies *IN*. (It is this possibility that scientists have in mind when they speak of "natural experiments.") In this respect, a manipulability theory that appeals to *IN* is quite different from traditional agency theories (such as those of von Wright

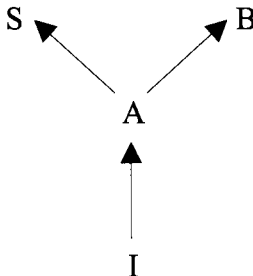


Figure 3.1.9



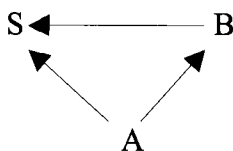


Figure 3.1.10

1971 and Menzies and Price 1993, discussed section in 3.4). In these theories, the characterization of a manipulation (or intervention) makes essential reference to human agency or free choice, and the hope is that this can be somehow grasped or understood independently of the notion of causality. By contrast, on the characterizations **IV** and **IN** there is nothing logically special about human action or agency: human interventions are regarded as events in the natural world like any other and they qualify or fail to qualify as interventions because of their causal characteristics and not in virtue of being (or failing to be) activities carried out by human beings.

**Circularity and Reduction** A third issue concerns circularity. The characterizations **IV** and **IN** employ causal language at a number of points: not only must the intervention variable  $I$  cause  $X$ , but  $I$  must not itself directly cause  $Y$ , must not be correlated with other causes of  $Y$  that are independent of the putative  $I \rightarrow X \rightarrow Y$  chain, and so on. Because the notion of an intervention is already a causal notion, it follows that one cannot use it to explain what it is for a relationship to be causal in terms of concepts that are themselves noncausal. Thus, a manipulability theory formulated in terms of **TC** or **M** and relying on the notion of an intervention characterized above will not allow us to translate causal claims into noncausal claims (whether the latter are framed in terms of, say, claims of statistical association of some sort or in terms of claims involving some noncausal notion of counterfactual dependence). However, it is also crucially important to understand that **IV** and **IN** are not viciously circular in the sense that the characterization of an intervention on  $X$  with respect to  $Y$  itself makes reference to the presence or absence of a causal relationship between  $X$  and  $Y$ . The causal information required to characterize the notion of intervention on  $X$  with respect to  $Y$  is information about the causal relationship between the intervention variable  $I$  and  $X$ , information about whether there

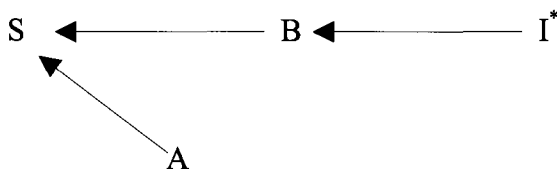


Figure 3.1.11

are other causes of  $Y$  that are correlated with  $I$ , information about whether there is a causal route from  $I$  to  $Y$  that does not go through  $X$  and so on, *but not information about the presence or absence of a causal relationship between  $X$  and  $Y$* . That there is a coherent notion of an intervention to be captured and that some explication of this notion that is not viciously circular must be possible are strongly suggested by the fact that we do seem to sometimes find out whether a causal relationship exists between  $X$  and  $Y$  by manipulating  $X$  in an appropriate way and determining whether there is a correlated change in  $Y$ . This fact by itself seems to show that we must have some notion of a manipulation of  $X$  that would be suitable for finding out whether  $X$  is causally linked to  $Y$ , and that this notion can be characterized without presupposing that there is a causal relationship between  $X$  and  $Y$ . It is just this notion that **IN** attempts to capture.

Because **IN** is not viciously circular, **TC** and **M**, which incorporate this characterization, are also not viciously circular in the sense that we already have to know whether there is a causal relationship between  $X$  and  $Y$  (or what its characteristics are) to apply them. The fundamental idea underlying **TC**, **M**, **SC**, and **NC\*** is that we can explain what it is for a relationship between  $X$  and  $Y$  to be causal by appealing to facts about *other* causal relationships (or the absence of such relationships) involving  $I$ ,  $X$ , and  $Y$  and to counterfactual claims concerning the behavior of  $Y$  under interventions on  $X$ . As explained above, these counterfactual claims are in turn understood in such a way that evidence regarding an association or correlation between changes in  $Y$  under interventions that change  $X$  are directly relevant to them. Thus, the pattern evinced by **SC**, for example, is this: *if a certain counterfactual is true (that  $Y$  or the probability of  $Y$  would change were an intervention to be carried out on  $X$ ), then a causal claim (that  $X$  causes  $Y$ ) will be true*. The antecedent of this counterfactual in turn makes reference to certain causal claims (because the notion of an intervention is a causal notion), but these are not claims about the presence or absence of a causal relationship between  $X$  and  $Y$  themselves. The existence of a correlation between  $X$  and  $Y$  under interventions on  $X$  is in turn evidence for the truth of this counterfactual. Thus, the overall pattern is this: facts about other causal relationships besides the relationship between  $X$  and  $Y$  + noncausal information about correlations  $\rightarrow X$  causes  $Y$ , where  $\rightarrow$  means "are evidence for."

Similarly, the pattern evinced by **NC\*** is this: certain causal facts (that  $X$  is a direct cause of  $Y$ ) imply that an existential claim in which a counterfactual is embedded is true (that there is some intervention on  $X$  such that if  $X$  is changed as a result of that intervention and all other causes of  $Y$  are held fixed at some value, then  $Y$  would change). The existence of a correlation between  $X$  and  $Y$  that persists under the interventions specified in the antecedent of this counterfactual is in turn evidence that the counterfactual is true. The overall pattern is that a causal fact of interest (that  $X$  is a direct cause of  $Y$ ) in conjunction with other causal facts (that  $X$  has been changed as a result of an intervention while other causes of  $Y$  are held fixed) provides a defeasible reason for expecting that a certain correlation will hold:  $X$  directly causes  $Y$  + other causal facts  $\rightarrow X$  and  $Y$  are correlated. Spelling out these patterns makes it

clear how a manipulability theory can provide a nontrivial *constraint* on what it is for a relationship to be causal without providing a reductive analysis of causality.

The fact that the relationships just described do not yield a reduction of causal claims to facts about correlations (or, for that matter, to claims about counterfactual dependence, where this notion is characterized without reference to causal notions) will be a source of disappointment to some philosophers. However, the failure of this sort of reduction has been a familiar theme in recent philosophical discussion (see, e.g., Cartwright 1983b) and, as I note below, there are systematic reasons why we should expect it. Moreover, there is a widespread consensus among both causal modelers and philosophers that reliable causal inference just on the basis of correlational evidence is not possible—not possible precisely because causal claims have a content (having to do with claims about the outcomes of hypothetical experiments) that is not reducible to correlational claims (see, e.g., Woodward 1995; Tufté 1974). Typically, a number of different causal structures will be compatible with a given body of correlational evidence, and to determine which of these structures is correct one must make use of additional background information, either domain-specific information about whether certain variables do or do not cause others, of the sort described in Cartwright (1979) and Woodward (1995), or domain-general principles, such as the Causal Markov condition connecting causal and correlational information or both. It is only given such additional background assumptions that there will be a connection between causal and correlational claims.<sup>3</sup> **TC** and **M** embody this point of view.

It is also worth noting that the nonreductive patterns associated with **TC** and **M** are common elsewhere in philosophy. Behaviorists hoped to translate claims about particular beliefs and desires into claims about overt behavior, but there is now a consensus that this cannot be done. The strongest such relationship that seems remotely plausible is something like this: given some psychological claim, that *S* desires *P*, and given various *other* psychological facts about *S*, facts about his other beliefs and desires, it follows that *S* will exhibit certain behavior. And given various other assumptions about *S*'s other beliefs and desires, certain kinds of behavior will be evidence that *S* desires *P*. In other words, any plausible connection between *S*'s psychological state and his overt behavior must make undischarged references to his other psychological states, and hence the behavioristic reduction fails. Similarly for the relationship between probabilities and relative frequencies.<sup>4</sup> Despite the failure of psychological claims to reduce to claims about overt behavior or probability claims to reduce to claims about relative frequencies, we can get a nontrivial purchase on what such claims mean by understanding how, given appropriate additional assumptions, they connect with facts about behavior or facts about frequencies. A parallel point holds in connection with the interrelations among causal, counterfactual, and correlational claims.

We may further illustrate these points about reduction and its connection to the manipulability theory by considering the causal structures in figures 3.1.12, 3.1.13, and 3.1.14. All of these systems are capable of generating the same pattern of correlations among *X*, *Y*, and *Z*. Indeed, even if we make the

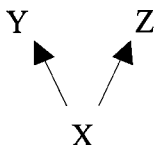


Figure 3.1.12



Figure 3.1.13

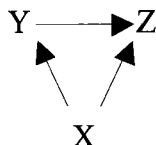


Figure 3.1.14

standard “screening off” assumptions linking causal structure to conditional independence embodied in the Causal Markov condition—that the effects of a common cause are independent conditional on the common cause, and so on—a researcher who simply observes the correlations  $X$ ,  $Y$ , and  $Z$  will be unable to determine which of figure 3.1.12 or figure 3.1.13 is correct, although she will be able to distinguish both from figure 3.1.14, as long as the structures generating these correlations remain stable or undisturbed. In both figures 3.1.12 and 3.1.13,  $Y$  and  $Z$  will be correlated and become independent conditional on  $X$ . In more complicated causal structures, the underdetermination problem often becomes much more massive. As I suggested above, from the perspective of a manipulationist account of causation, underdetermination is a reflection of the fact that the causal claims embodied in figures 3.1.12, 3.1.13, and 3.1.14 have additional content over and above the claim that certain correlations obtain. While figures 3.1.12 and 3.1.13 agree about which correlations and conditional correlations obtain, they disagree about what would happen under certain interventions. For example, 3.1.12 claims that there is an intervention on  $X$  that will change both  $Y$  and  $Z$  but that an intervention on  $Y$  will not change  $Z$  or  $X$ . By contrast, 3.1.13 claims that there is an intervention on  $X$  that will change  $Z$  but no intervention on  $X$  will change  $Y$  and that there is an intervention on  $Y$  that will change both  $X$  and  $Z$ . Of course, *if* these interventions should happen to occur, then different patterns of correlations will actually be generated by figures 3.1.12 and 3.1.13. However, though this *may* happen—for example, if there is an experimenter around with the desire to intervene and the right technology for doing so, or if nature should just happen to produce the right sort of natural experiment—on a manipulationist theory nothing guarantees that it *must* happen. In other words, on a manipulationist account, the extent to which differences in causal structure will be reflected in (or translatable into claims about) differences in patterns of correlations is an empirical question, the answer to which will turn on whether various possibilities (interventions or other sorts of occurrences that, modulo certain assumptions, are equivalent to interventions) happen to occur. There is no conceptual guarantee that such translation always must be possible.

### 3.1.6 Comparison with Other Notions of Intervention

I remarked above that there are a number of characterizations of the notion of an intervention (or at least of closely related notions) in the philosophical and nonphilosophical literature, including Spirtes et al. ([1993] 2000), Pearl (1995,

2000a), Hausman (1998), and Cartwright and Jones (1991). These characterizations are broadly similar to **IN** in the sense that they attempt to capture the notion of an ideal experimental manipulation but differ in detail. I will not attempt a comprehensive comparison, but several points are worth noting. First, **IN** is similar in a number of respects to the characterization that Spirtes et al. (pp. 75ff) give of a “manipulation.” But one important point of difference is that their characterization assumes that the directed graph representing the manipulation variable(s) and the system in which the manipulation occurs come with an associated probability distribution that satisfies the Causal Markov condition **CM**. By contrast, **IN** does not assume that the system intervened on satisfies **CM**.<sup>5</sup> This difference is in turn linked to the use that Spirtes et al. put their notion to: they are mainly interested in inferring or calculating the results of manipulations from information about graphical structure and an accompanying probability distribution. For this purpose, **CM** is extremely useful. My aim, by contrast, is to provide an account of the content of various causal notions, and for this purpose it seems more suitable to adopt a characterization of “intervention” that does not assume **CM**.

A second issue concerns another important choice that we face concerning the characterization of interventions. Consider a spring that (within a certain range of extensions) conforms to Hooke’s law  $F = -kX$ , and imagine a manipulation of the extension that satisfies the conditions for an intervention in **IN** but stretches the spring so much that it breaks. One possible view taken by, among others, Cartwright and Jones (1991) and perhaps Pearl (2000a) is that such a manipulation should not count as a bona fide intervention; in general, manipulations of  $X$  that alter or destroy the mechanism by which  $X$  affects  $Y$  should not count as “admissible ways of changing  $X$  for the purposes of determining whether  $X$  causes  $Y$ .” These writers suggest that we should build into **IN** the requirement that an intervention on  $X$  should not change the causal mechanism or relationship (if any) connecting  $X$  to  $Y$ . Call this the mechanism-preserving requirement. Obviously, **IN**, as formulated above, does not impose such a requirement.

The motivation for the mechanism-preserving requirement may seem obvious: if our manipulation of  $X$  destroys the causal mechanism that connects  $X$  to  $Y$ , so that  $Y$  does not change under manipulation of  $X$ , then we may be misled into thinking that there is no causal relationship between  $X$  and  $Y$ , when in fact such a relationship exists.<sup>6</sup> Note, though, that we will make this mistaken inference only if we formulate the connection between causation and manipulation as a claim about what will happen under “all” rather than, as **TC** and **M** do, under “some” interventions. If we formulate the connection in terms of “some” interventions, we will not be justified in concluding that  $X$  does not cause  $Y$  just because there is some intervention on  $X$  (a manipulation that destroys the mechanism connecting  $X$  to  $Y$ ) that does not change  $Y$ . Instead, we may conclude that, for example, extending the spring causes it (in the total and contributing cause sense) to exert a restoring force as long as it is true (as it clearly is) that some interventions that change  $X$  will change  $F$ .

One apparent consequence of adopting the mechanism-preserving requirement is that to determine whether we have carried out an intervention on  $X$ , we must have some basis for determining whether our manipulation of  $X$  has disrupted the mechanism, if any, connecting  $X$  to  $Y$ . This in turn seems to require that we already have some information about the causal relationship, if any, between  $X$  and  $Y$  and introduces a worry about a kind of “circularity” that seems to be potentially much more vicious than the circularity built into **IN**. While **IN** builds information about other causal relationships, besides the relationship between  $X$  and  $Y$ , into the characterization of an intervention on  $X$ , the mechanism-preserving requirement builds into that characterization information about the very thing we want to characterize: the causal relationship between  $X$  and  $Y$ .

Despite this concern, I do not think the circularity associated with the mechanism-preserving requirement is always or automatically vicious. For one thing, one can *sometimes* recognize that a contemplated manipulation of  $X$  is likely to disrupt any causal relationship between  $X$  and  $Y$ , should any exist, without knowing whether there is in fact such a relationship.<sup>7</sup> Nonetheless, it is also true that we sometimes find out about whether there is a causal relationship between  $X$  and  $Y$  and about its characteristics by means of relatively “black box” experiments—by manipulating  $X$  in circumstances in which we have little, if any, prior information about this causal relationship. This in turn suggests that there must be some legitimate characterization of the notion of an intervention on  $X$  that builds in little or no information of this sort. **IN** attempts to provide such a characterization. In part for this reason, in part because it is often unclear how to determine whether an intervention has disrupted a mechanism, and in part because, in view of my discussion in the preceding paragraph, there is no overriding reason for adopting the mechanism-preserving requirement, I prefer to avoid building the requirement into the characterization of interventions.

Adopting **IN** has the additional virtue of permitting us to say (as I have been saying throughout this essay) that extensions that break the spring can be genuine interventions; it is just that the generalization connecting  $X$  to  $F$  is not invariant or does not continue to hold under such interventions, although it does hold under others. More generally, given **IN**, we may talk about a generalization being invariant under some range of interventions and breaking down under others. By contrast, if we adopt the mechanism-preserving requirement, it will not make sense to talk of a generalization’s breaking down under some interventions but not others: if the generalization fails to hold for some manipulation of the value of its independent variables, this will show that the manipulation in question was not really an intervention.

Having said this, I will add that I suspect (although I will not try to show) that many of the claims about the connection between causation and intervention in this essay can also be expressed in a framework in which the preferred notion of intervention conforms to the mechanism-preserving requirement, as long as that requirement is understood in a way that allows that one can

sometimes determine whether a manipulation on  $X$  is mechanism-preserving without already knowing whether  $X$  causes  $Y$ .

Cartwright and Jones (1991) impose the mechanism-preserving requirement *in addition to* a set of requirements that are similar to those in **IN**. By way of contrast, Pearl (2000a) characterizes what he calls an “atomic” intervention as follows:

**(PI)** The simplest type of external intervention is one in which a single variable, say  $X_i$ , is forced to take on some fixed value  $x_i$ . Such an intervention, which we call “atomic,” amounts to lifting  $X_i$  from the influence of the old functional mechanism  $x_i = f_i(pa_i, u_i)$  and placing it under the influence of a new mechanism that sets the value  $x_i$  while leaving all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by  $do(X_i = x_i)$  or  $do(x_i)$  for short, amounts to removing the equation  $x_i = f_i(pa_i, u_i)$  from the model and substituting  $X_i = x_i$  in the remaining equations. (p. 70)

The requirement that an intervention on  $X$  break the connection between  $X$  and its parents is paralleled by I1 and I2 in **IN**, but there is no obvious counterpart to I3 and I4 in **PI**. In addition, of course, **PI** differs from **IN** in requiring that *all* other mechanisms, including the mechanism, if any, connecting  $X$  to its effects, be preserved under an intervention on  $X$ . Can we make do with a notion of intervention that just involves **PI** and nothing else?

Although Pearl’s notion of an intervention is ideally suited to the main use to which he puts it—that of predicting what will happen to the probability distribution of one variable under a certain kind of change in the value of another (or others) when the causal graph (or some portion of it) connecting those variables and all others in the system of interest is known—it seems unsuitable for the semantic project pursued in this essay. (This is not intended as a criticism of Pearl, who has quite different purposes.) One way of seeing this is to note that **PI** defines the notion of an intervention with respect to the *correct* causal graph for the system in which the intervention occurs, because it is this graph that specifies what the causal relationships are that must be left “unperturbed” by the intervention. Because of this, **PI** does not seem to give us a notion of intervention that can be used to provide an interpretation for what it is for such a graph to be correct. Put slightly differently, the difficulty is that **PI** takes the notion of a causal mechanism (including the notion of a mechanism linking  $X$  to  $Y$ ) as primitive, and defines the notion of an intervention in terms of it, thus losing any possibility of using the latter notion to define the former.

There is another limitation of **PI** that may be brought out by means of the following example.<sup>8</sup> Incubated children at risk for retrolental fibroplasia were injected with vitamin E. It turned out that “the actual effective treatment was opening the pressurized oxygen-saturated incubators several times a day to give the injections, thus lowering the barometric pressure and oxygen levels in the blood of the infants” (Pearl 2001). The injection of vitamin E appears to constitute an intervention in the sense of **PI**: the injection disrupts whatever mechanism previously controlled the children’s vitamin E levels and does not,

at least in any obvious way, disrupt other (relevant) mechanisms. But although the incidence of retrolental fibroplasias is lower for children who are injected than for children who are not, the injections do not cause this decrease and the decrease is not a “causal effect” (direct or otherwise) of the injections.

In a discussion of this experiment, Pearl (2001) writes: “In your example of the vitamin E injection (above), there is another variable being manipulated together with  $X$ , namely the incubator cover,  $Z$ , which turns the experiment into a  $do(x,z)$  condition instead of  $do(x)$ . Thus, the experiment was far from ideal, and far even from the standard experimental protocol, which requires the use of placebo.” One limitation of this response is that it does not really give us any insight into what was wrong with the original experiment or why a placebo should have been used. Any experiment, no matter how ideal, will involve the manipulation of more than one variable. Even in a well-designed experiment with a proper placebo and control (the incubators for the children in the control group are opened and they are injected with some inert substance), the opening of the door will change the values of many other variables (the position of air molecules, etc.) in both the treatment and control group. For this reason, one cannot capture what was defective in the original experiment by observing that it involved the manipulation of two variables  $X$  and  $Z$  rather than just  $X$ . Instead, what was wrong in the original experiment has to do with the *relationship* between the additional variable  $Z$  (besides  $X$ ) that is manipulated and the outcome of interest: incidence of retrolental fibroplasia ( $R$ ). In particular, the problem is not just that  $Z$  as well as  $X$  is changed, but rather that  $Z$  affects  $R$  independently of  $X$ . If, as in the example involving the air molecules given above, the manipulation affected a second variable  $A$  as well as  $X$  but there was no path from  $A$  to  $R$  that fails to go through  $X$ , there would be (at least in this respect) nothing wrong with the experiment. The conditions I3 and I4 capture this feature of good experimental design. If our interest is in characterizing a notion of intervention that can be used to capture the content of causal claims, we need to include these conditions.<sup>9</sup>

### 3.1.7 Causal Relata

Yet another issue raised by **IN** concerns the relata of the causal relationship. It should be clear, both from **IN** and our preceding discussion, that carrying out an intervention on  $X$  requires that, for whatever unit or entity is characterized by  $X$ , there be a well-defined notion of changing the value of  $X$  for that unit in such a way that the very same unit is caused by the intervention to possess a different value of  $X$ . For example, an intervention that consists of administering a drug to some subject must change whether that very subject has been given the drug. As explained above, this in turn requires that it be possible in some sense for the subject to be in either of (at least) two different states: to be either treated with the drug or not treated. For similar reasons, the effect variable  $Y$  must be such that it is possible for the same subject to assume more than one value of this variable: either to recover or not to recover.



It is standard in philosophy to think of the relata in type-causal claims as properties or event types. However, a manipulability account suggests a rather different perspective. A manipulability account implies that for something to be a cause we must be able to say what it would be like to change or manipulate it. This in turn suggests (as we have been assuming) that within a manipulability framework it is most natural or perspicuous to think of causes and effects not as properties, but as *variables*, or more precisely, as *changes in the values of variables*, where one of the characteristics of a variable is that it is capable of taking two or more values and of being changed from one of these values to another. On this way of looking at matters, causal claims tell us not that one property is associated with or necessitates another, but rather that certain changes in the value of a variable will produce associated changes in the value of another variable.

As we have already noted, in some cases, it will be straightforward and unproblematic to translate talk of causal relationships between properties or event types into variable talk, by making use of so-called indicator variables, taking the values 1 or 0, to express the occurrence or nonoccurrence of the properties or event types in question. In such cases, the advantage of the formulation in terms of variables is that it is more transparent and precise. We made use of this sort of formulation at the beginning of this section, when we expressed the claim that treatment with drug *D* causes recovery in terms of a claim about the relationship between changes in the value of an indicator variable *T* measuring whether or not treatment with *D* was administered and a second indicator variable *R*, measuring whether or not recovery occurs. Similarly, in chapter 2, section 4, rather than talking of (2.4.3) the event type or property of being hit with a hammer causing the event type or property of shattering, we introduced an indicator variable *H* that takes the values {*hit*, *not hit*} or {1, 0}, depending on whether some object of interest is hit with momentum greater than  $m_1$  or not hit at all, and a variable *S* that takes the values {*shatter*, *not shatter*}, depending on whether that object shatters. Representing the causal relationship between being hit and shattering as a relationship between *H* and *S* had the virtue of making it transparent that what we are really claiming, when we assert (2.4.3), is that a certain change in the value of *H* (from *not hit* to *hit*) will produce a corresponding change in the value of *S* (from *unshattered* to *shattered*). Although this is somewhat less transparent on the property formulation (2.4.3), it is natural to think of the variable formulation as simply capturing in a somewhat clearer way what was intended all along by the property formulation.

There are other cases, however, in which claims about causal relationships between properties are not readily understandable as claims about relationships between variables because the whole idea of changing those properties does not seem to be well-defined. For example, there are properties that, for logical, conceptual, or metaphysical reasons, must be possessed by every object—properties that, so to speak, can only take one value (that of being present). These properties will fail to satisfy **IN**, and hence there will be no well-defined notion of an intervention with respect to them. For example, if it

is metaphysically necessary that everything that exists is a physical object or if we lack any coherent conception of what it is for something to exist but to be non-physical, then there is no well-defined notion of intervening to change whether something is a physical object. Although there are true (and even lawful) generalizations about all physical objects, in a manipulability theory these will not describe causal relationships. For example, although presumably it is a law of nature that (L) no physical object can travel at a velocity greater than light, (L) is not, according to a manipulability theory, a *causal* generalization. It would be a mistake to interpret (L) as the claim that being physical causes objects to move at a velocity less than or equal to that of light, for this would be to claim that manipulating whether something is physical is a way of changing whether or not it moves faster than light.

Moreover, even with respect to variables that can take more than one value, the notion of an intervention will not be well-defined if there is no well-defined notion of changing the values of that variable. Suppose that we introduce a variable "animal" which takes the values {*lizard, kitten, raven*}. By construction, this variable has more than one value, but if, as I suspect, we have no coherent idea of what it is to change a raven into lizard or kitten, there will be no well-defined notion of an intervention for this variable. Of course, we might keep a raven in a cage and replace it with a lizard or a kitten, but this is not to change one of these animals into another. What is changed in this case is the contents of the cage, not the animals themselves. Thus, the relevant variable is not the variable "animal," but a variable that takes as its values various possible contents of the cage. Again, though there may be true or even lawful generalizations about kinds of animals, these will not be causal generalizations.<sup>10</sup>

This restriction on the notion of an intervention to variables for which there is a well-defined notion of change is both implicit in the notion of an intervention itself and also follows from our guiding idea that causal relations are relations that can be used for manipulation and control. If there is no well-defined notion of changing the value of *X*, we cannot, even in principle, manipulate some other variable by changing *X*.

It would be desirable to have some general and precise characterization of what it is for a change in the value of a variable to be "well-defined," but I am unable to supply this. In particular cases, the reader will often have an intuitive understanding of what this notion involves, and in what follows I rely on this. Even in the absence of a general characterization, it is interesting and important that the notion of causation is so closely bound up with the notion of changing the value of a variable.

One consequence of these considerations is that a number of properties or conditions that are often thought to be causes are at best problematic candidates for this role; examples include the property of being a member of a certain species, being a member of a particular race, and being a certain age. In each case, the notion of an intervention that changes the values of these "variables" does not appear to be well-defined. It might be thought that it is an arbitrary stipulation to claim that such properties cannot be causes, but in fact, as we

shall see (section 3.2), causal claims involving them are genuinely unclear precisely because it is unclear what hypothetical experiments to associate with them. In such cases, we can often sharpen or clarify the meaning of causal claims involving unmanipulable properties by replacing them with claims involving variables, the values of which are manipulable, and specifying more precisely just what it is that they imply about the outcomes of hypothetical experiments. In other cases, the problem is not so much that there is no interpretation of the cause variable under which it is manipulable, but that there are a number of different possible interpretations and associated with these a number of different possible hypothetical experiments. Here too, being explicit about just which hypothetical experiment is the intended interpretation of a causal claim can play an important role in clarifying its meaning. Again, this is just what we should expect if some version of the manipulability theory is correct.

### 3.1.8 *IN as a Regulative Ideal*

I emphasized above that the point of the conditions in **IN** is not that reliable causal inference is possible only when those conditions are met. Instead, the role of **IN** (and thus of **M** and **TC**) is to serve as a kind of regulative ideal: they tell us what must be true of the relationship between *X* and *Y* if *X* causes *Y* and in this way tell us what we should aim at establishing, perhaps on the basis of nonexperimental evidence or on the basis of an imperfect or nonideal experiment, if we want to show that a causal claim is true. In other words, when we engage in causal inference regarding the effects of *X* in a situation in which there is no variable that satisfies all of the conditions for an intervention variable with respect to *X* (or at least there is no such variable that we are able to observe or measure), we should think of ourselves as trying to determine what would happen in an ideal hypothetical experiment in which *X* is manipulated in such a way that the conditions in **IN** are satisfied. For example, in those cases in which we want to know what the effect of *X* on *Y* is but must rely on nonexperimental data in which some independent cause *Z* of *Y* is correlated with *X* and we try to correct for this confounding influence, what we are trying to do is determine how *Y* would change if, contrary to actual fact, *X* was changed by intervention *I* (where *Z* is not correlated with *I* and the other conditions in **IN** are met as well). We will return to this point about the regulative role of **IN** below, when we consider the sense in which interventions must be “possible.”

## 3.2 Causal Claims and Hypothetical Experiments

I maintained above that a manipulability theory suggests that one can clarify the meaning of causal claims by spelling out the hypothetical experiments with which they are associated. To the extent that the hypothetical experiment associated with a causal claim is left unclear—either because the relevant

manipulation is not well-defined or because what is being claimed about what would happen under this manipulation is left unspecified—the causal claim itself will be unclear or ambiguous. In such cases, we often can clarify or disambiguate causal claims by explicitly distinguishing among different possible claims about the outcomes of hypothetical experiments that might be intended or by replacing claims involving factors for which the notion of manipulation is not well-defined with claims involving variables for which there is a well-defined notion of manipulation.<sup>11</sup> The fact that we can clarify the meaning of causal claims in this way is in turn an additional reason for accepting a manipulability account of causation.

As an illustration of this idea, consider the following causal claim:

(3.2.1) Being female causes one to be discriminated against in hiring and/or salary.

Claims of this sort are quite common, but I contend, following Holland (1986), that they are fundamentally unclear. The problem is not so much that under all interpretations of the putative cause (“being female”) we lack any clear idea of what it would be like to manipulate it, but that (a) there are several rather different things that might be meant by manipulation of “being female” (which is to say that there are several quite different variables we might have in mind when we talk about being female as a cause), and the consequences for salary discrimination of manipulating each of these may be quite different. Moreover, (b) the manipulations one is most likely to have in mind when one asserts (3.2.1)—what one probably intends by (3.2.1)—do not involve literal manipulations of gender at all.

One claim that might be intended by (3.2.1) is this:

(3.2.2) Employers determine whether to hire *A* and how much to pay *A* at least in part on the basis of their beliefs about *A*'s gender, independently of what else they believe about *A*'s qualifications, credentials, background, and so on; that is, differences in employer beliefs about the gender of applicants cause differences in whether applicants are offered a job or the salary they are offered.

Associated with this causal claim is the following claim about the outcome of a hypothetical experiment: if one were to intervene to change an employer's belief that a candidate is male (female) to a belief that the candidate is female (male), the result would be to decrease (increase) the probability that the candidate is hired and decrease (increase) the salary that the candidate is offered.

In cases in which candidates are hired sight unseen just on the basis of some written record, the relevant manipulation may be easy to carry out; it might consist in altering information on a candidate's application about his or her gender while leaving other information intact: changing male to female and vice versa on application forms and suppressing other information that contains clues as to gender. In other cases in which candidates have to appear for an interview, such direct manipulation of an employer's beliefs may be

difficult or impossible, although one might envision other related experiments that might serve as fallible and imperfect vehicles for getting at what would happen under a belief manipulation, given that one cannot actually carry it out; for example, one might consider sending pairs of candidates to job interviews, one male and one female, but with qualifications otherwise matched as closely as possible, and observing the results. These are all perfectly straightforward, well-defined, and nonmysterious kinds of manipulations, and if it turns out that employers are more likely to hire or to offer higher salaries to candidates who they believe to be males than to equally qualified and experienced candidates who they believe to be females, this would provide strong support that, when interpreted this way (i.e., as (3.2.2)), (3.2.1) is correct. Notice, though, that the variable (or cause) that is manipulated in these experiments is not really gender, but the employer's beliefs about gender; what (3.2.2) amounts to is the claim that manipulation of employer beliefs will change an applicant's salary and probability of hiring.

We need to distinguish such belief manipulations, and the interpretation (3.2.2) associated with them, from interventions that would actually change a candidate's gender and from the associated interpretation of (3.2.1) as a claim about what would happen to hiring and salary under such interventions. Unlike an "intervention" that changes a raven into a kitten, a gender-changing intervention is conceptually well-defined and physically possible. In fact, there are a number of different forms such an intervention might take, some of which human beings can at present carry out and others of which may be possible at some point in the future. These range from a manipulation of a subject's genotype shortly after conception that replaces an X with a Y chromosome or vice versa to "massive doses of hormones in utero that would lead to female morphology at birth" to a "sex change operation" at some point before or after birth (cf. Rubin 1986, p. 962).

It seems plausible that these different sex-changing interventions may have different effects on a subject's hiring and salary. It seems even more plausible that at least some of these will also have different consequences from the interventions that change an employer's beliefs considered in the preceding paragraph. Suppose that if the belief-manipulating experiments had been carried out in a particular industry, subjects who are believed to be male would have been observed to have a 25 percent higher probability of being hired and a 15 percent higher wage than females. There is, I take it, no particular reason to believe that we would have observed exactly the same differentials if instead we had carried out the various hypothetical manipulations of applicant gender described above and then allowed them to apply for jobs. The fact that some range of manipulations of employer *beliefs* about gender will change salary (i.e., that the relationship between employer beliefs about gender and salary is invariant under some range of interventions that change those beliefs) tells us little about which, if any, manipulations of *gender* will change salary. In spelling out what (3.2.1) means, we thus need to make it clear exactly what interpretation is intended. In fact, it seems clear that

those who assert (3.2.1) rarely, if ever, have in mind this second possibility, in which gender is actually physically manipulated.

There is yet another possible interpretation of (3.2.1); in fact, in a discussion of this example, the statistician Paul Holland (1996) claims that it is the preferred interpretation of (3.2.1) or what one ought to mean by (3.2.1):

In the salary discrimination example it is easy to get confused as to what the counterfactual of interest really is. For example, it is *not* what a woman's salary would have been had she been a man with the same qualifications, even though it is the coefficient on the gender variable that attracts all the interest (and we even talk about "holding qualifications constant")! Rather we want to know what the salaries of men and women would be if there were no discrimination... In the salary discrimination example the data are all collected in an allegedly discriminatory system so we have *no* data relevant to what the salaries would have been under a system of non-discriminatory salary administration. (p. 56)

Here, the suggestion is that we should not think of the manipulated variables as either gender or beliefs about gender, but rather as the existence of a certain set of employment practices, which may be either discriminatory or nondiscriminatory: it is discriminatory employment practices (or to put it more perspicuously, the contrast between discriminatory and nondiscriminatory practices), and not gender per se, that is claimed to be the cause of salary differentials. An intervention on this variable presumably would involve legal, institutional, and cultural changes that would alter the present discriminatory regime to a nondiscriminatory regime or vice versa. On this interpretation, one thinks of (3.2.1) as a claim about what would happen to women's salaries under such a change; presumably, they would be higher under the nondiscriminatory regime and the amount by which they would be higher is a measure of the extent of salary discrimination women face.

I see no reason to follow Holland in thinking that this is the only legitimate interpretation of (3.2.1), but I fully agree that it is *a* plausible interpretation, one that captures what is sometimes meant by the assertion of (3.2.1), and furthermore, one that commits us to a quite different counterfactual (a claim about the outcome of a quite different hypothetical experiment) and hence a quite different causal claim than the two previous interpretations. (It is presumably also the interpretation that is most directly relevant to policy.) Holland is also almost certainly right in thinking that, because a change to a nondiscriminatory regime of the sort he envisions would involve massive shifts in institutions, attitudes, and behavior, it is problematic to use data gathered from either of the first two (belief-changing and gender-changing) experiments described above, which takes place entirely in a discriminatory regime to predict what would happen to salaries under a nondiscriminatory regime. Here again, we see that it matters a great deal to the interpretation and truth conditions of (3.2.1) which hypothetical manipulation we have in mind.

### 3.3 Realism about Causation

I turn now to some remarks on the relationship between a manipulability account of causation and "realism" about causation. Many philosophers have advocated accounts of causation that are antirealist or subjectivist in the sense that they represent causal relationships as built out of two distinct components: a component that is objectively "out there" in nature (this usually has to do with the existence of certain regularities or correlations, perhaps supplemented with facts about spatiotemporal relationships) and a component that is in some way made up or added by us, having to do with facts about our psychology or mental organization (e.g., our expectations that certain regularities and not others will persist, or our inclination to organize our experience in certain ways and not others). The core idea of such accounts is that the difference between those regularities that reflect causal relationships and those that do not is not an intrinsic, objective difference, but has to do with a difference in our beliefs or attitudes regarding these two classes of regularities.

As we have already noted, both philosophers who have advocated manipulability theories of causation and their critics have thought that such theories lead directly to a conception of causation that is subjectivist in this sense. For example, Menzies and Price (1993) take their version of a manipulability theory to imply that causation is a "projection" onto the world of our experience of human agency and that causation is thus a "secondary quality" like color. They argue that claims about causal relationships cannot be understood without reference to the standpoint human beings adopt when they act as agents, just as (they suppose) "red" must be understood by reference to a characteristic experience that human perceivers have.

What does the version of the manipulability theory I advocate say about such issues? Subjectivists are not very forthcoming about just what is meant by their claims that causation is "projected" onto the world by us or "constituted" by our beliefs and attitudes. On one straightforward and literal reading of these contentions, it follows that the truth values of causal claims are in some way dependent on the existence of human beings or their beliefs, attitudes, or experiences: if human beings did not exist or had different beliefs, attitudes, or experiences, then the truth values of causal claims would be different or they would lack truth values. I suggested above that there may be a limited respect in which this is true: which causal claims we accept as true (or at least *readily* accept) are influenced by what we take to be a "serious possibility." However, once we fix which possibilities are serious, it seems to me that, putting aside those causal claims that are overtly about the effects of human psychological states (as when a patient's recovery from a disease is influenced by whether she has an optimistic attitude), there is no further sense in which the counterfactuals about the outcomes of the hypothetical experiments associated with typical causal claims are in some way dependent on human attitudes or beliefs. In other words, the counterfactuals on which causal claims are based seem to be true or false in a mind-independent way, even if it is true that the causal claims themselves reflect additional assumptions

about which possibilities are serious. Consider, for example, the hypothetical experiment in which I step in front of a speeding bus. Whether I will be injured in such an experiment does not depend, either causally or in some other way, on my beliefs or desires. Similarly, there is an obvious intuitive sense in which it is facts about how the world is and not facts about my expectations or projective activities that determine what will happen to my longevity in the experiment in which I purchase life insurance considered in section 2.1.

Indeed, it seems very hard to make sense of the activity of conducting experiments to assess the correctness of causal claims if the truth of those claims is somehow partly dependent on or constituted by the experimenter's beliefs or expectations. If the "objective" core of the content of the claim that *X* causes *Y* is just the claim that *X* and *Y* are correlated and all else is the product of some agent's projective activities, what sense can we make of experiments designed to distinguish the claim that *X* causes *Y* from the claim that they are correlated because of the operation of some common cause? Are such experiments simply roundabout ways of finding out about the experimenter's (or the scientific community's) projective activities? Within a subjectivist framework, what is the rationale for such familiar features of experimental design as taking care to ensure that the experimental manipulation of *X* is not correlated with other causes of *Y*? Of course, it is true that my beliefs and expectations will influence my *beliefs* about what the outcome of these experiments will be, but this isn't to say that they will influence the outcome of the experiments themselves were they to be carried out. To suppose otherwise would be to attribute a kind of magical power to beliefs and expectations that they plainly don't possess.<sup>12</sup>

We may drive this point home by focusing again on the structure of the notions of intervention and invariance. Consider an agent who wishes to change *Y* but cannot change it "directly," who can change *X* directly, and who wonders whether she can change *Y* by changing *X*. Assume again that the case is not one that involves psychological causation in an ordinary, unproblematic sense. Then, although there will be a sense in which it may be up to the agent (and hence dependent on her beliefs and attitudes) whether she chooses to bring about *X*, it is a presupposition of her deliberation that it is not also up to her whether, if *X* occurs, *Y* will occur—whether *X* is or is not a means to *Y*. Instead, it is a presupposition of her deliberation that if it is possible to change *Y* by intervening on *X*, then there must be an independently existing, invariant relationship between *X* and *Y* that the agent makes use of when she changes *X* and, in doing so, changes *Y*—a relationship that would exist and have whatever characteristics it has even if the agent were unable to manipulate *X* or chose not to manipulate *X* or did not exist. In other words, it is built into the whole notion of a manipulation that the agent's activities, manipulative or otherwise, don't somehow create or influence or constitute whether there is a relationship between *X* and *Y* that allows us to manipulate *Y* by manipulating *X*. Thus, when I deliberate about the claim that if I were to step in front of a speeding bus, I would be seriously injured,



I assume that the truth value of this claim is independent of my choices, deliberative activities, and state of mind. If this were not the case, it would be hard to make sense of my deliberations. Contrary to what many philosophers have supposed, a commitment to some version of realism about causation (in the sense that relationships of counterfactual dependency concerning what will happen under interventions are mind-independent) seems to be built into any plausible version of a manipulability theory.

In this view of the matter, putting aside the point that assessments of serious possibility may be influenced by an agent's interests, there is no further sense in which causal claims are mind-dependent. Rather than somehow serving as a basis out of which causal relationships are created or constructed (as traditional versions of manipulability theories claim), our practical interests as agents serve, so to speak, to pick out the kind of (independently existing) relationship between *X* and *Y* that we are interested in when we worry about whether that relation is causal.<sup>13</sup> What the relationships we label causal have in common (and the respect in which they contrast with merely correlational relationships) is that they support potential manipulations in the way described above.

To put the point slightly differently: on the view I am advocating, our notion of causality developed in response to the fact that there are situations in which we could manipulate *X*, and by so doing manipulate *Y*. This fact led us (3.3.1) to form the notion of a relationship between *X* and *Y* that would support such manipulations and to contrast this with the notion of a mere correlation that would not support such manipulations. However, it is built into the notion of a relationship that will support manipulations in this way that (3.3.2) such a relationship would continue to hold even if we do not or cannot manipulate *X*, or if our beliefs and attitudes were different, or even if we did not exist at all. If it is asked why (3.3.2) is built into our notion of causation, my response is that any other view of the matter would involve a bizarre and magical way of thinking, according to which our ability to manipulate *X* or our practical interest in manipulating *X* or our beliefs about the results of manipulating *X* somehow make it the case that a means-end connection comes into existence between *X* and *Y* where this connection would not exist if we did not have the ability or interest or beliefs in question. Taken literally, such a view, if intelligible at all, would require human beings to have god-like powers that they plainly do not possess.

This conclusion is reinforced by the naturalistic, evolutionary perspective endorsed in chapter 2. According to subjectivist accounts, causal relationships have their source in facts about us—facts about our expectations, attitudes, and so on—which we “project” on to the world. If we think about this claim in the light of the argument of section 2.1, the subjectivist picture looks rather peculiar. To begin with, what is the evolutionary story about the benefits we derive from this projective activity? After all, our projectivist tendencies systematically lead to beliefs that, by the subjectivist's own account, are mistaken or ungrounded—mistaken in the sense that they ascribe a false objectivity to causal claims or involve thinking of the distinction between causal and

correlational claims as having an objective basis in nature rather than in facts about us. Why should we and other animals go to the trouble of distinguishing between causal and correlational relationships if all that is “really out there” in the world are correlations? All that projecting seems wasteful and gratuitous.<sup>14</sup> Moreover, why do our projective activities take the particular form they do? In some cases, a single co-occurrence is sufficient for belief in a causal connection, whereas in other cases, repeated co-occurrences have no such effect. Why is this? Saying that what distinguishes causal relationships from mere correlations are facts about our attitudes and projective activities gives us no insight into why our attitudes are such that they lead us to interpret a single episode of mushroom consumption followed by nausea as indication that the former causes the latter, but do not lead us to make a similar inference when confronted with evidence of an extensive correlation between purchase of life insurance and increased longevity.

By contrast, an objectivist account of causation fits much more readily with the empirical observations discussed in section 2.1. If there is an objective difference between those correlations that reflect relationships that can be used for purposes of manipulation and control and those that cannot, and if sensitivity to this difference can have a practical payoff in terms of an increased probability of survival and reproduction, then there is no mystery about why the ability to learn how to discriminate between such relationships should have evolved. Of course, organisms that are suitably cognitively complex will have attitudes, expectations, and beliefs that lead them to regard certain sequences but not others as causal, but they have these attitudes, beliefs, and expectations precisely because their possession facilitates reliable discrimination between two kinds of sequences that differ objectively. In other words, rather than, as the subjectivist supposes, the organism’s beliefs, attitudes, and expectations somehow creating or constituting the difference between causal and merely correlational relationships, it is the prior, independent existence of an objective distinction between cause and correlation (and the fact that sensitivity to this difference may have important fitness consequences for the organism) that explains why organisms have the different beliefs, attitudes, and expectations regarding causal and noncausal sequences that they do. We are willing to infer, on the basis of a single experience, that the relationship between consumption of a certain sort of mushroom and subsequent nausea is causal precisely because it is often or at least sometimes true that nausea is caused by (and not merely correlated with) what we eat, and there are great practical advantages to recognizing the difference between a causal and a merely correlational relationship in this sort of case.

I emphasize that the kind of realism that follows from this way of viewing matters is metaphysically modest and noncommittal. It requires only that there be facts of the matter, independent of facts about human abilities and psychology, about which counterfactual claims about the outcome of hypothetical experiments are true or false and about whether a correlation between *C* and *E* reflects a causal relationship between *C* and *E* or not. Beyond this, it commits us to no particular metaphysical picture of the “truth makers”

for causal claims. Nonetheless, it suggests that the implications of a manipulability theory for the mind-independence of causation are very different from what is ordinarily assumed.

I conclude this section with some additional remarks about counterfactuals. Like the notion of causation itself, counterfactuals have often been regarded with suspicion. It is frequently suggested that they lack a clear meaning or that their truth conditions are so vague and context-dependent that they are not suitable for understanding or elucidating any notion (of causation or anything else) that might be of scientific interest. A famous example of Quine's illustrates the worry. Consider the counterfactual(s)

(3.3.3) If Julius Caesar had been in charge of U.N. Forces during the Korean War, then he would have used (a) nuclear weapons or (b) catapults.

It is hard to see on what basis one could decide whether the counterfactual (3.3.3) with (a) as consequent or the counterfactual (3.3.3) with (b) as consequent (or neither) is correct.

A manipulability framework for understanding causation provides a response to this worry. It suggests that the appropriate counterfactuals for elucidating causal claims are not just any counterfactuals, but rather counterfactuals of a very special sort: those that have to do with the outcomes of hypothetical interventions. As argued above, it does seem plausible that counterfactuals that we do not know how to interpret as (or associate with) claims about the outcomes of well-defined interventions will often lack a clear meaning or truth value. For example, (3.3.3a) and (3.3.3b) seem unclear for just this reason. It isn't just that we lack the technological means to carry out an experimental manipulation in which Caesar is placed in charge of the U.N. Forces. The more fundamental problem is that we have no clear conception of what would be involved in carrying out such an experiment.

By contrast, a similar sort of skepticism about counterfactuals that are interpretable as claims about the outcomes of hypothetical (but otherwise well-specified) interventions is much harder to sustain. Consider again the hypothetical experiment described in section 3.1, in which a large group of people suffering from a disease are randomly divided into a treatment and a control group, with the former receiving some drug that is withheld from the latter, and the right sort of experimental controls being followed. If this experiment were actually carried out and the incidence of recovery was much higher in the treatment group than in the control group, it would be natural to think of it as providing good evidence for the truth of counterfactuals like the following:

(3.3.4) If those in the control group had received the drug, the incidence (or expected incidence) of recovery in that group would have been much higher.

In the case in which this experiment is actually carried out, the claim that (3.3.4) lacks a determinate meaning or truth value is much less plausible than

the corresponding claim about (3.3.3). Moreover, assuming that the experiment is well-designed, whether or not the experiment is actually carried out does not determine whether or not (3.3.4) has a truth value or what that truth value is. We think instead of (3.3.4) as having a determinate meaning and truth value whether or not the experiment is actually carried out—it is precisely because the experimenters want to *discover* whether (3.3.4) is true or false that they conduct the experiment. Of course, the researchers may be mistaken in the conclusion they draw about the truth value of (3.3.4), but this does not distinguish (3.3.4) from any other empirical knowledge claim.

### 3.4 Agency Theories of Causation

We may further clarify the account of the relation between causation and manipulation that I favor by contrasting it with the manipulability theory developed by Menzies and Price (1993). These writers claim, in contrast to the position that I have defended, that one can understand or grasp notions like agency and manipulation by a human agent independently of the notion of causation and that one can then use these notions to provide a noncircular “analysis” of causation.

I argue that this theory is subject to the following set of dialectical difficulties. To show that the notion of agency is independent of or prior to the notion of causality, one needs to give human actions or manipulations a special status: these can't be ordinary causal transactions. This in turn has two problematic consequences. First, it flies in the face of any plausible version of naturalism: it makes agency out to be a fundamental, irreducible feature of the world and not just one variety of causal transaction among others. Second, it leads us toward an undesirable kind of anthropomorphism or subjectivism regarding causation, just as critics of traditional manipulability theories have charged. If the only way we understand causation is by means of our prior grasp of the experience (or notion) of agency, then we face an obvious problem about the extension of causal notions to circumstances in which manipulation by human beings is not possible, for these will be circumstances in which the relevant experience of agency is unavailable.

Menzies and Price's (1993) basic thesis is that “an event *A* is a cause of a distinct event *B* just in case bringing about the occurrence of *A* would be an effective means by which a free agent could bring about the occurrence of *B*” (p. 187). They take this connection between free agency and causation to support a probabilistic analysis of causation (according to which “*A* causes *B*” can be plausibly identified with “*A* raises the probability of *B*”), provided that the probabilities appealed to are what they call “agent probabilities,” where “agent probabilities are to be thought of as conditional probabilities, assessed from the agent's perspective under the supposition that the antecedent condition is realized *ab initio*, as a free act of the agent concerned. Thus the agent probability that one should ascribe to *B* conditional on *A* . . . is the probability that *B* would hold were one to choose to realize *A*” (p. 190).

As with IN, Menzies and Price recognize that the key feature of this characterization is that it involves the postulation of a special kind of independent, exogenous causal history for the putative cause *A*: genuine causal relationships between *A* and *B* are those that will persist under changes in *A* brought about in this special way, whereas spurious relations between *A* and *B* will not. However, unlike the view I have defended, Menzies and Price attempt to appeal to the notion of agency to provide a noncircular, reductive analysis of causation. They claim that circularity is avoided because we have a grasp of the *experience* of agency that is independent of our grasp of the general notion of causation:

The basic premise is that from an early age, we all have direct experience of acting as agents. That is, we have direct experience not merely of the Humean succession of events in the external world, but of a very special class of such successions: those in which the earlier event is an action of our own, performed in circumstances in which we both desire the later event, and believe that it is more probable given the act in question than it would be otherwise. To put it more simply, we all have direct personal experience of doing one thing and thence achieving another. We might say that the notion of causation thus arises not, as Hume has it, from our experience of mere *succession*; but rather from our experience of *success*; success in the ordinary business of achieving our ends by acting in one way rather than another. It is this common and commonplace experience that licenses what amounts to an ostensive definition of the notion of “bringing about.” In other words, these cases provide direct non-linguistic acquaintance with the concept of bringing about an event; acquaintance which does not depend on prior acquisition of any causal notion. An agency theory thus escapes the threat of circularity. (pp. 194–95)

However, Menzies and Price also recognize that, once the notion of causation has been tied in this way to our “personal experience of doing one thing and hence achieving another” (p. 194), a problem arises concerning unmanipulable causes. To use their own example, what can it mean to say that “the 1989 San Francisco earthquake was caused by friction between continental plates” (p. 195) if no one has (or, given the present state of human capabilities, could have) the direct personal experience of bringing about an earthquake by manipulating these plates? Their response to this difficulty is complex, but the central idea is captured in the following passages:

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

Clearly, the agency account, so weakened, allows us to make causal claims about unmanipulable events such as the claim that the 1989 San Francisco earthquake was caused by friction between continental plates. We can make such causal claims because we believe that there is another situation that models the circumstances surrounding the earthquake in the essential respects and does support a means-end relation between an appropriate pair of events. The paradigm example of such a situation would be that created by seismologists in their artificial simulations of the movement of continental plates. (p. 197)

The problem with this suggestion becomes apparent when we consider, for example, the nature of the “intrinsic” but (allegedly) “noncausal” features in virtue of which the movements of the continental plates “resemble” the artificial models the seismologists are able to manipulate. It is well-known that small-scale models and simulations of naturally occurring phenomena that superficially resemble or mimic those phenomena may nonetheless fail to capture their causally relevant features because, for example, the models fail to “scale up”—because causal processes that are not represented in the model become quite important at the length scales that characterize the naturally occurring phenomena. Thus, when we ask what it is for a model or simulation that contains manipulable causes to “resemble” phenomena involving unmanipulable causes, the relevant notion of resemblance seems to require that the same *causal* processes are operative in both. I see no reason to believe (and Menzies and Price provide no argument) that this notion of resemblance can be characterized in noncausal terms. But if the extension of their account to unmanipulable causes requires a notion of resemblance that is already causal in character and that, *ex hypothesi*, cannot be explained in terms of our experience of agency, then their reduction fails.

More generally, Menzies and Price are subject to the following dilemma. On the one hand, if their references in the passages quoted above to intrinsic features that either “support a means-end relation” or are “identical with (or closely similar to) intrinsic features” that are present in “an analogous pair of means-end related events” are meant to suggest that these features are fully reducible to facts having to do with our experience of agency and facts about noncausal relationships of similarity to situations in which manipulable causes and the experience of agency are present, then they have provided no reason to think that such a reduction can be carried out. On the other hand, the passages may be meant to suggest that, quite independently of our experience or perspective as agents, there is a certain kind of relationship with intrinsic features that we exploit or make use of when we bring about *B* by bringing about *A*. Moreover, because this relationship is intrinsic and can exist independently of anyone’s experience of agency, it can also be present even when *A* is not in fact manipulable by humans. If so, I would claim that this is essentially the objectivist position regarding the connection between causality and agency that I have endorsed: considerations having to do with agency and manipulability help to explain why we developed a notion of causality having the features it does and play a heuristic role in helping to

characterize the meaning of causal claims, and have considerable epistemic relevance when we come to test causal claims, but agency is not in any way “constitutive” of causality. This view yields a far more plausible treatment of causes that are not manipulable by human agents and avoids the problems that result from taking agency to be a primitive feature of the world, but it also abandons any pretense of noncircular reduction of causality to agency.

There are two other features of Menzies and Price’s proposal that are worth comment by way of distinguishing their position from my own. First, Menzies and Price’s view of the origins of our concept of causality is a thoroughly empiricist one: we derive or learn the concept entirely from a characteristic kind of experience. As they see it, what is wrong with Hume’s account is simply that he fixes on the wrong candidate for the relevant experience: it is our experience of acting as agents rather than the experience of regular succession that is crucial. But, as should be clear from chapter 2, the idea that our concept of causation is derived purely from experience (whether of agency or of anything else) is simply mistaken. As with other concepts, the acquisition of the concept of causality involves a complicated interaction between prespecified neural mechanisms and “learning.” Moreover, only some forms of learning are based on “experience” in the sense of that notion that Menzies and Price have in mind.<sup>15</sup> Our practical interests as agents shapes the operation of our brains and the learning strategies we employ and hence our notion of causality, but this is not to say that we have acquired the concept of causality just from the experience of being agents. There is no reason why a theory that takes the connection between causation and agency seriously should also be committed to the empiricist picture of concept acquisition advocated by Menzies and Price. This point is of considerable importance because it is this picture of concept acquisition that helps to ground the reductive features of their project.

The second point has to do with Menzies and Price’s claim that the correlations involving *A* that will persist under the supposition that *A* “is realized ab initio as a free act” give us the effects of *A*. Menzies and Price do not further explain what they mean by the quoted phrase, preferring, as we have noted, to indicate what they mean by ostension, by pointing to our experience as agents. It should be clear from the characterization of an intervention in section 3.1, however, that whether (as soft determinists would have it) free action is understood as an action that is uncoerced or unconstrained or due to voluntary choices of the agent, or whether, as libertarians would have it, a free action is an action that is uncaused or not deterministically caused, the persistence of a correlation between *A* and *B* when *A* is realized as a free action is neither necessary nor sufficient for *A* to be the cause of some effect. Suppose that *A* is realized via a free act (in either of the above senses) and remains correlated with *B* when produced in this way, but that *A* is also correlated with *C*, another cause of *B*. (This possibility is compatible with *A*’s being free in either of the above two senses.) Then, as we have seen, it need not be true that *A* causes *B*. Suppose, then, that we respond to this difficulty by adding to our

characterization of *A*'s being a free act the idea that *A* must not itself be correlated with any other cause of *B*. (Passages in Price 1991 suggest such an additional proviso, although the condition in question seems to have nothing to do with the usual understanding of free action.) As we have seen, even with this proviso, it need not be the case that *A* causes *B* if *A* remains correlated with *B* when *A* is produced by an act that is free in this sense, because it still remains possible that the free act that produces *A* also directly causes *B*. (This would be the case, to use an example discussed in section 3.3, in which the experimenter's administration of a drug to a treatment group has a direct effect on recovery via a placebo.)

Menzies and Price's underlying idea that *A* causes *B* if the association between *A* and *B* persists when *A* is given the right sort of independent causal history is correct, but the relevant notion of an independent causal history is a rather complex notion, given by the characterization of an intervention **IN** and not just by the notion of a free action. To be sure, free actions often do have, or can be made to have, the characteristics **IN**, but the fact that they are free doesn't guarantee that this is so.

For similar reasons, it is not necessary that *A* be realized by a free human action for it to have the right sort of causal history for the application of the test under discussion. An action that causes *A* but is unfree (because it is caused or caused in the wrong sort of way) can qualify as an intervention or as a suitable history of *A* for the purposes of assessing its effects on *B*, as long as it satisfies the conditions **IN**. Indeed, as we have seen, causal processes that produce changes in *A* but involve no human actions at all can qualify as interventions as long as they meet these conditions. In general, in determining whether the fact that the association between *A* and *B* persists under some process that produces changes in *A* shows that *A* causes *B*, it is the causal characteristics of the process that produce these changes in *A*, as described in **IN**, that matter, not whether a free human action figures in the process.

If this is correct, it further undermines the reductionist aspect of Menzies and Price's project. Whatever the merits of the idea that we have a grasp of the notion of agency, via direct experience, that is prior to or independent of our grasp of the general notion of causality, we surely have no corresponding direct experiential grasp of what it is for our interventions to satisfy the conditions **IN**. Indeed, it seems plain that to grasp what it is for our actions to satisfy these conditions, we must have a prior grasp of the notion of causality, and much more besides.

### 3.5 In What Sense Must Interventions Be Possible?

I return now to an issue raised in chapter 2. In what sense must an intervention on *X* be "possible" if *X* is to cause *Y*?

It should be clear, in view of our discussion so far, that the relevant notion of possibility has nothing to do with what human beings can do. Because the notion of an intervention can be characterized without reference to human



activities, it makes sense to speak of hypothetical interventions on  $X$ , of what would happen to  $Y$  under such interventions, and hence of  $X$  causing  $Y$ , even if manipulation of  $X$  is not within the technological abilities of human beings, and indeed even in circumstances in which human beings or other agents do not exist. In worlds in which human beings do not exist, natural processes with the right causal characteristics will qualify as interventions. On the other hand, we also noted that if we cannot think of  $X$  as a variable that is capable of being changed from one value to a different value—if manipulation of  $X$  is not a logical possibility or if it is ill-defined for conceptual or metaphysical reasons—then claims about what will happen to  $Y$  under interventions on  $X$  will either be false or will lack a clear meaning. This gives us one respect in which interventions on  $X$  must be “possible” if  $X$  is to cause  $Y$ : interventions on  $X$  must at least be logically possible and well-defined.

Must interventions on  $X$  be physically (i.e., nomologically) possible for  $X$  to cause  $Y$ ? That is, if  $X$  causes  $Y$ , must it be physically possible for a causal process to occur that qualifies as an intervention on  $X$ ? In thinking about this problem, it will be useful to distinguish two different notions of physical possibility. On one notion, an event  $E$  is physically possible if and only if it is consistent with the laws of nature and the actually obtaining initial conditions. When conjoined with determinism, this notion of physical possibility implies that interventions on  $X$  will not be possible unless they actually occur. For example, if the experimenters in the drug experiment described in section 3.1 do not in fact intervene to give the drug to some subject, this intervention will be “impossible.” In view of our discussion above, it should be obvious that it would be much too strong to require that interventions on  $X$  must always be possible in this sense, for, given **TC** and **M**, this would be to make the existence of a causal connection between  $X$  and  $Y$  dependent on whether an intervention on  $X$  actually occurs.

A considerably weaker notion of physical possibility is the following:  $E$  is physically possible if and only if there is some set of possible initial conditions (“possible” in the sense that the conditions themselves are consistent with the laws of nature), perhaps different from those that actually obtain, such that the occurrence of  $E$  is consistent with those conditions and the laws of nature. In other words,  $E$  is physically impossible if and only if its occurrence is ruled out by the laws of nature alone, independently of facts about initial conditions. In this sense, it is presumably physically possible, even if determinism is true, for the experimenters to withhold treatment from a subject even though they did not do so. It may be that withholding would have occurred only if initial conditions at some earlier time had been different but, given these different initial conditions, no further violation of physical law would be required for the withholding to occur. Thus, this weak notion of possibility gives us a sense in which it is possible for subjects who did not receive the treatment to have received it and vice versa. By contrast, an intervention that required the acceleration of a particle from a velocity less than that of light to a velocity greater than that of light would be physically impossible even in this weaker sense.

It is of considerable interest that there appear to be cases in which  $X$  causes  $Y$  but interventions on  $X$  are not physically possible even in the weak sense. Consider the (presumably) true claim

(3.5.1) Changes in the position of the moon with respect to the earth and corresponding changes in the gravitational attraction exerted by the moon on various points on the earth's surface cause changes in the motion of the tides.

On a manipulability theory, this claim implies that in some hypothetical experiment in which the gravitational attraction exerted by the moon is varied by varying its distance from the earth, the motion of the tides would change. Let us now ask whether an intervention that changes the distance between the earth and the moon is physically possible in the weak sense described above. On this notion, when we ask, for example, whether there is a physically possible intervention that would double the moon's orbit, we are asking whether there is some physically possible process (involving the occurrence at some earlier time of initial conditions that are different from the actual initial conditions prevailing at that time but involving no violation of physical law) that leads from the actual world to a situation in which the moon's orbit is twice its present value in a way that satisfies the conditions for an intervention (**IN**). Here the constraint of physical possibility means, for example, that we are not allowed to imagine the moon moving from its present position to a new position at a superluminal velocity or for the moon to change its orbit except under the influence of some impressed force of appropriate direction and magnitude, where the force in question must obey the usual source laws and general laws of motion such as  $F = ma$ .

It is not at all clear that there is such a physical process. All physically possible processes that would change the position of the moon might be, in Elliott Sober's words (personal correspondence), "too ham-fisted" to satisfy all of the conditions in **IN**. For example, one way to change the position of the moon would be to change the position of some other massive body in such a way that it exerts a gravitational force on the moon that changes its position. However, a change in the position of this massive body would exert a direct gravitational effect on the tides, in violation of condition (I3) in **IN**. Similarly, we might imagine that the position of the moon is changed by means of a collision with some other body. However, not only would this body have a direct gravitational influence on the motion of the tides, but the impact itself would presumably change the shape and mass distribution of the moon and thus the gravitational force the moon exerts on the tides. The resulting change in the tides would reflect not just the effects of the change in the position of the moon alone, which is what we are interested in, but the change in the moon's mass distribution. Perhaps with sufficient ingenuity we may be able to think up some physically possible intervention that changes the position of the moon and that satisfies the conditions **IN**, but I hope to have made it plausible that nothing in the truth of the original causal claim (3.5.1) *guarantees* that this will be possible. What we need for such an intervention is a physically possible

process that is sufficiently fine-grained and surgical that it does not have any other effects on the tides besides those that occur through the change that it produces in the position of the moon, and it may well be that the laws of nature guarantee that all real causal processes will have such additional effects. At the very least, it seems wildly optimistic to assume that appropriately surgical intervention processes must be available for all true causal claims.

What is at issue here may be made clearer by means of a second, less realistic example. Suppose that

(3.5.2) Cs only occur spontaneously in the sense that they themselves have no causes.

In other words, there are no further factors  $C^*$  that affect whether or not  $C$  happens, and this is a matter of physical law. (I take this to be at least a logically coherent possibility.) Thus, it is physically impossible to carry out an intervention that changes whether  $C$  occurs. Nonetheless, it seems quite possible that Cs themselves might well have further effects  $E$ .

This example brings out that to demand that if  $C$  is to cause  $E$ , it must be physically possible to intervene on  $C$ , is to demand that if  $C$  is to have effects, it must itself be such that it can be affected by some other cause. This demand in turn seems to conflict with the apparently reasonable idea that whether  $C$  causes  $E$  ought to turn just on the nature of the relation between  $C$  and  $E$ , and not on whether  $C$  itself is caused (or could be). More generally, to make whether (3.5.3)  $C$  causes  $E$  depend on whether interventions on  $C$  are physically possible seems to be to make (3.5.3) depend on what sorts of causal histories are possible for  $C$  itself, and this consideration seems extrinsic and irrelevant to the nature of the relation between  $C$  and  $E$ .

At this point, it will be helpful to return to an observation made earlier: that the notion of an intervention characterized by **IN** represents a regulative ideal. Its function is to characterize the notion of an ideal experimental manipulation and in this way to give us a purchase on what we mean or are trying to establish when we claim that  $X$  causes  $Y$ . We have already noted that for this purpose, it isn't necessary that an intervention actually be carried out on  $X$ . All that is required is that we have some sort of basis for assessing the truth of claims about what would happen if an intervention *were* carried out. Similarly, I suggest that as long as there is some basis for assessing the truth of counterfactual claims concerning what would happen if various interventions were to occur, it doesn't matter that it may not be physically possible for those interventions to occur.

This may become clearer if we return to our original motivation for introducing the notion of an intervention. This was twofold. First, we wanted to exclude cases in which we had no coherent conception of what it was to change the variable intervened on. Second, we wanted to exclude cases involving confounding—cases in which, although an association between  $C$  and  $E$  persists under changes in  $C$ , this is due to something other than a direct causal link from  $C$  to  $E$ , such as an independent cause of  $E$  that is correlated with  $C$  or a direct effect of the intervention itself on  $E$ . The notion of an

intervention was designed to ensure that if the changes in *C* had a certain kind of history, these sorts of possibilities would not arise. This suggests that there will be a basis for claims about what will happen to *E* under an intervention on *C* as long as we can associate some well-defined notion of change with *C* and as long as we have some grounds for saying what the effect, if any, on *E* would be of changing just *C* and nothing else. This means, in particular, that there must be a way of disentangling—perhaps merely conceptually or analytically rather than in actuality—the effect on *E* of changing just *C* from the effects on *E* of changes in other potentially confounding variables, including direct effects from the intervention process itself.

In each of the examples described above, these conditions are met. In the case of (3.5.1), Newtonian gravitational theory and mechanics themselves provide the needed basis. Although it may be true that any actual physical process that changes the position of the moon will also directly influence the tides, Newtonian theory and familiar rules about the composition of forces tell us how to subtract out any direct influence from such a process so that we can calculate just what the effect of, say, doubling of the moon's orbit (and no other changes) would be on the tides, even though it also may be true that there is no way of actually realizing this effect alone. In other words, Newtonian theory itself delivers a determinate answer to questions about what would happen to the tides under an intervention that doubles the moon's orbit, and this is enough for counterfactual claims about what would happen under such interventions to be legitimate and to allow us to assess their truth.

In the drug experiment, what is of interest are counterfactual claims about what would happen to various subjects if they were to either receive or not receive the drug. As we have seen, there is a sense in which, under determinism, it is "impossible" for any subject who did not receive the drug to receive it, but this impossibility does not undercut the legitimacy of the counterfactual claim, made of some subject who received the drug and recovered, that if he had not received the drug he would not have recovered. This counterfactual makes sense because we have a clear notion of what it is to change the status of a subject from someone who receives the drug to someone who does not (and vice versa) and because we have a clear conception of what it would be to remove or think away the influence of confounding influences other than the effect of the drug alone on recovery. The design of the experiment, which involves assessing the above counterfactual by looking at the disease status of other subjects who did not receive the drug in a randomized experiment, reflects this understanding.

A similar conclusion seems appropriate in connection with (3.5.2). First, there is nothing about the example that suggests that the notion of a change in *C* is logically impossible or not well-defined. Second, there is nothing about the structure of the example that undercuts the possibility of separating out the effect on *E* of just a change in *C* from the effect on *E* of changes in other variables. On the contrary, the structure of the example itself seems to provide a basis for this. To begin with, if there are no possible causal processes that affect *C*, we don't have to worry about the possibility of any such process

directly affecting  $E$  in violation of (I3). Moreover, if we are willing to accept the principle of the common cause, then, given that the occurrence of  $C$  is genuinely spontaneous, it cannot be correlated with any independent cause of  $E$  in violation of (I4). Even if, in violation of the principle of the common cause, there are other, independent causes of  $E$  that “naturally” occur in such a way that they are correlated with  $C$ , we may be able to interfere with their occurrence in such a way as to make them uncorrelated with  $C$ , or alternatively, we may be able to observe their influence on  $E$  when  $C$  is absent, and use this as a basis for subtracting out their influence in cases in which  $C$  is present. These considerations give us a purchase on what the counterfactual “If  $C$  were to be changed by an intervention,  $E$  would change” means (and how we might in principle determine whether it is true or false), even in circumstances in which there is no way of intervening on  $C$ . If we find, as the description of the example suggests, a persistent correlation between  $C$  and  $E$  in such circumstances, this is reason to accept the counterfactual.

This way of looking at matters supports the conclusion we reached above on intuitive grounds: that it would be a mistake to make the physical possibility of an intervention on  $C$  constitutive in any way of what it is for there to be a causal connection between  $C$  and  $E$ . This would be to repeat, in a somewhat more general way, the mistake made by agency theorists, who claim that there can be no causal relation between  $C$  and  $E$  unless manipulation of  $C$  is within human capabilities. When an intervention changes  $C$  and in this way changes  $E$ , this exploits an independently existing causal link between  $C$  and  $E$ . One can perfectly well have the link without the physical possibility of an intervention on  $C$ .

My conclusion, then, is that at least in circumstances like those described in the above examples, we may meaningfully make use of counterfactual claims about what would happen under interventions, even when such interventions are not physically possible in either the strong or weak senses described above, and we can legitimately use such counterfactuals to elucidate causal claims along the lines suggested by **M** and **TC**. In other words, the reference to “possible” interventions in **M** and **TC** does not mean “physically possible”; instead, an intervention on  $X$  with respect to  $Y$  will be “possible” as long as it is logically or conceptually possible for a process meeting the conditions for an intervention on  $X$  with respect to  $Y$  to occur. The sorts of counterfactuals that cannot be legitimately used to elucidate the meaning of causal claims will be those for which we cannot coherently describe what it would be like for the relevant intervention to occur at all or for which there is no conceivable basis for assessing claims about what would happen under such interventions because we have no basis for disentangling, even conceptually, the effects of changing the cause variable alone from the effects of other sorts of changes that accompany changes in the cause variable.

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

causal claims, there are no physically possible processes that are sufficiently fine-grained or surgical to qualify as interventions.

These observations are relevant in an obvious way to David Lewis's counterfactual theory of causation (1973, [1973] 1986b, [1979] 1986c, 2000). One of the features of Lewis's theory that has been most resisted by commentators is his insistence that in many cases, the right counterfactuals to consider in explicating causal claims will be those whose antecedents require the occurrence of a "miracle," a violation of some law. Although my version of a manipulability theory differs in various respects from Lewis's theory (see section 3.6), it agrees with his on this important point.

### 3.6 Comparison with Lewis's Theory

As I have indicated, the version of the manipulability theory for which I have been arguing has a number of affinities with the counterfactual theory of causation developed by David Lewis. This section explores the relationship between Lewis's theory and my own in more detail. I focus on the deterministic version of the theory in Lewis (1973, [1979] 1986c, 1986b). I believe that the observations made in this section apply equally well to Lewis's (2000) most recent version of a counterfactual account, but I do not attempt to discuss this account in this chapter. (The notion of "influence," which figures prominently in Lewis 2000, is discussed briefly in chapter 5.) The most obvious point of similarity between Lewis's theory and the account that I have been defending is that both approaches assign a central role to counterfactuals in elucidating causal notions; both thus stand in opposition to accounts like Salmon's (1994), which attempt to avoid reliance on counterfactuals. I assume that the basic features of Lewis's theory are sufficiently well-known that a detailed summary is not required. Unlike the account I have been defending, Lewis's theory is intended only to be an account of token causation between particular events. Causation is defined as the ancestral of counterfactual dependence:  $c$  causes  $e$  if and only if  $c$  and  $e$  occur and there is a chain of counterfactual dependence between  $c$  and  $e$ , that is, a sequence of events  $c_1, c_2 \dots c_n$  such that  $e$  is counterfactually dependent on  $c_n$ ,  $c_n$  is counterfactually dependent on  $c_{n-1} \dots$ , and  $c_1$  is counterfactually dependent on  $c$ . If we let  $O(c)$ ,  $O(e)$  be the propositions that the events  $c$  and  $e$  occur, then  $e$  is counterfactually dependent on  $c$  if and only if the following counterfactuals hold:

$$(3.6.1) \quad O(c) \rightarrow O(e)$$

$$(3.6.2) \quad \neg O(c) \rightarrow \neg O(e)$$

If  $c$  and  $e$  both occur (as must be the case if  $c$  causes  $e$ ), then, on Lewis's semantics, the counterfactual (3.6.1) is automatically true. Hence, what is crucial on Lewis's theory for whether  $e$  is counterfactually dependent on  $c$  is whether the counterfactual (3.6.2) is true.

For Lewis, a counterfactual claim of form “if  $p$  were the case, then  $q$  would be the case” is true if and only if there is a possible world in which both  $p$  and  $q$  are true that is “more similar” or “closer” to the actual world than any possible world in which  $p$  is true and  $q$  false. The criteria for evaluating “similarity” among possible worlds that are appropriate for the sorts of counterfactuals used to elucidate causal claims are as follows:

- (S1) It is of the first importance to avoid big, widespread, diverse violations of law.
- (S2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
- (S3) It is of the third importance to avoid even small, localized simple violations of law.
- (S4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (1986, p. 47)

As Lewis explains (1986, appendix B, p. 56), “big, widespread, diverse violations of law” are events that “consist of many little miracles together, preferably not all alike. What makes a big miracle more of a miracle is not that it breaks more laws, but that it is divisible into many and varied parts, any one of which is on a par with a little miracle.”

As I noted above, one important difference between Lewis’s theory and my own is that the former is an account of causation between particular events and is not intended to cover type-causal claims. By contrast, my account focuses primarily on type-level claims and applies to token-causal claims only insofar as these may be viewed as instances of type claims in the sense captured by **AC\***. Nonetheless, I assume in what follows that we may abstract away from this difference and ask whether Lewis’s theory and the criteria (S1–4) for assessing similarity can be extended to type-causal claims, and how the results of doing so compare with an intervention-based approach. We may also, of course, ask how Lewis’s theory compares with the account of token causation **AC\***. To raise only the most obvious possibility, if it were true that when we extend the criteria (S1–4) to the evaluation of counterfactuals relating types of events, or when we compare Lewis’s theory to **AC\***, we are led to the same results as intervention-based approaches, this would both be interesting in its own right and would also raise the question of whether the interventionist account gains us anything that is not already available from Lewis. To avoid cumbersome qualifications, I speak in what follows of how Lewis’s theory compares to the interventionist account, but the reader should bear in mind that what is really being compared is either an extension of Lewis’s theory to type causation and **M** (or **TC**), or else Lewis’s theory of token causation and **AC\***.

Let me begin with an example designed to illustrate the broad similarities between Lewis’s approach and my own. Consider the counterfactual

- (3.6.3) If  $e_1$  had not occurred, then  $e_2$  would not have occurred

evaluated with reference to a deterministic causal structure in which  $c$  is the common cause of two joint effects,  $e_1$  and  $e_2$ , neither of which causes the other. Because counterfactual dependence is sufficient for causation in Lewis's theory, we want this counterfactual to come out false. Lewis's criteria achieve this result because the most similar world to the actual world is a world (world 1) that matches the actual world exactly up to just before  $e_1$  occurs, at which point, a small localized miracle occurs that results in the nonoccurrence of  $e_1$ . In this world, both  $c$  and  $e_2$  will still occur, and hence the counterfactual (3.6.3) will come out false. By contrast, in a world (world 2) in which the nonoccurrence of  $e_1$  is achieved through the earlier nonoccurrence of  $c$  as the result of a small miracle just before the time  $c$  would have occurred, we still require a miracle, and there is a less extensive region of perfect match of particular fact with the actual world than is the case with world 1, because divergence from the actual world begins earlier, with the nonoccurrence of  $c$ . Accounting for the nonoccurrence of  $e_1$  by introducing a miracle at some still earlier time would produce an even less extensive region of perfect match. In this way, Lewis's theory arrives at the result that so-called backtracking counterfactuals such as

(3.6.4) If  $e_1$  had not occurred, then  $c$  would not have occurred

are false, and moreover, that counterfactuals like

(3.6.5) If  $e_1$  had not occurred, then  $c$  still would have occurred

are, as one intuitively expects, true.

The description of this example should make it clear that Lewis's similarity criteria and, in particular, the "small localized miracles" they require often function in broadly the same way as the notion of an intervention.<sup>16</sup> Again abstracting from the type/token difference, on both approaches, when we evaluate a counterfactual of the form "if  $C$  had not occurred, then  $E$  would not have occurred" with respect to a world in which  $C$  does occur, we think of the antecedent of this counterfactual as made true by some exogenous source of change—an intervention or a localized miracle—which breaks whatever endogenous causal relationships are at work in the actual world in producing  $C$ . This gives the nonoccurrence of  $C$  an independent causal history, and if the miracle has the right character (in particular, if it has no other effects on  $E$  except those that occur through the nonoccurrence of  $C$ ), and if it is inserted in the right place (see below for the significance of both of these qualifications), it follows that any change in  $E$  will be an effect of just the change from a situation in which  $C$  occurs to one in which it does not, and not the result of a change in some other factor. In other words, the invocation of "miracles" works, to the extent that it does, because, like the notion of an intervention, it requires us to consider a counterfactual situation in which what changes is only whether  $C$  occurs. Nothing else changes that might have an effect on  $E$  independently of  $C$ , thus ensuring that if  $E$  does change, this can only be because it is an effect of  $C$ . In opposition to those (e.g., Bennett 1984; Hausman 1998) who claim that a theory of counterfactuals should countenance only a single interpretation of



counterfactuals that permits backtracking and is appropriate for both causal and noncausal contexts, the manipulationist account agrees with Lewis in holding that there is a fundamental distinction between those similarity criteria that are appropriate for the counterfactuals that may be used to analyze causal claims and that should not permit backtracking, and those criteria that are appropriate for backtracking counterfactuals. Moreover, as we have already noted in section 3.5, demanding that changes in  $C$  be due to an intervention often has the consequence that they must involve “miracles” or violations of laws of nature, just as Lewis claims. The manipulationist account agrees with this feature of Lewis’s theory as well.

Despite these similarities, there are deep differences between Lewis’s theory and the version of the manipulability theory that I have been defending. One of these concerns the question of reduction. Lewis’s theory is avowedly reductionist in aspiration: the idea is to define the notion of causation in terms of a more general notion of counterfactual dependence that does not itself presuppose causal notions. By contrast, as explained above, the manipulationist account does not purport to provide such a reduction. According to the manipulationist account, given that  $C$  causes  $E$ , which counterfactual claims involving  $C$  and  $E$  are true will always depend on which other *causal* claims involving other variables besides  $C$  and  $E$  are true in the situation under discussion. For example, it will depend on whether other causes of  $E$  besides  $C$  are present. Thus, in figure 2.3.1 with  $a = -bc$ , the causal relationship between  $X$  and  $Y$  fails to reveal itself in a straightforward Lewisian pattern of counterfactual dependence of  $Y$  on  $X$  because of the presence of another cause  $Z$  of  $Y$ . To reveal the direct causal dependence of  $Y$  on  $X$  we must invoke a more complex counterfactual than any that figures in Lewis’s account: a counterfactual about how the value of  $Y$  would change if an intervention were to hold  $Z$  fixed and if at the same time another intervention were to change the value of  $X$ . This sort of counterfactual, with its reference to *two* interventions, one of which is on the putative cause  $X$  but the other of which is on variable  $Z$ , which is off the direct route from  $X$  to  $Y$ , has no direct analogue in Lewis’s system.<sup>17</sup> Moreover, if there are several indirect routes between  $X$  and  $Y$ , as in figure 2.3.4, then a counterfactual with an even more complex antecedent in which variables on all of the indirect routes between  $X$  and  $Y$  are fixed is required. Thus, which counterfactuals are appropriate for capturing the counterfactual dependence of  $Y$  on  $X$  when  $X$  is a direct cause of  $Y$  will depend on the causal features, including the causal route structure, of the larger system in which  $X$  and  $Y$  are embedded: the specification of which additional variables should be held fixed in the antecedent of the counterfactual relating  $X$  to  $Y$  requires reference to causal facts about the directed path structure of the larger system in which  $X$  and  $Y$  figure. The manipulationist account does not assume that (or try to show how) reference to such causal facts can be eliminated in favor of purely noncausal counterfactual claims. This point of view is also embodied in the characterization **IN** of an intervention and in **TC** and **M**, both of which, in assessing or elucidating the claim that  $X$  causes  $Y$ , make explicit reference to the truth of various other

causal claims and make no attempt to discharge these references in noncausal terms.<sup>18</sup>

We may think of Lewis as attempting to achieve the same results through the use of the similarity criteria (S1–4) that the manipulability theory achieves through reliance on explicitly causal notions. Lewis intends these to be understood in a way that is noncausal or that does not presuppose causal ideas. Hence, they are a suitable basis for reducing causation to counterfactual dependence. I argue below that this strategy is only partially successful. To the extent that the similarity criteria are clear, they often lead to the same results as the manipulationist account, but not always; sometimes, Lewis's rules lead us to insert miracles in the "wrong" place and generate mistaken evaluations of counterfactuals and hence of causal claims, assuming the connection between counterfactual dependence and causation that Lewis advocates. In such cases, the intervention-based approach is superior. I argue that these limitations in Lewis's theory derive from its reductionist aspirations. What they show is that we need the distinctively causal features of the notion of an intervention and counterfactuals whose antecedents make reference to causal information in order to formulate a satisfactory version of a counterfactual theory of causation. A noncausal similarity metric such as (S1–4) is an inadequate surrogate for these.

Another important respect in which the intervention-based theory differs from Lewis's is in its motivation. Lewis is explicit that his standards for similarity are adopted simply because they make those counterfactuals come out true that we think, pretheoretically, ought to be true. The standards themselves are complex and not particularly intuitive. As several writers (including, in particular, Horwich 1987, pp. 171ff) note, this leaves an important question unanswered: Why should we have this particular set of similarity criteria? What is the larger point or rationale that lies behind our use of these standards? For example, why don't we employ a set of standards in which S3 is weighted more heavily than S2, or in which, in contrast to S4, some rather than little or no weight is attached to approximate similarity in matters of particular fact? To respond that the standards S1–4 are preferred because they are the standards that are reflected in "our" notion of causation simply invites queries about why that notion is so special. Why should we not (why, in fact, did we not) develop a notion of "smausation" instead, connected to counterfactuals in the way that "causation" is in Lewis's theory, but according to which counterfactuals are evaluated by some different set of similarity criteria?

This issue becomes even more pressing when we consider that, as emphasized above, small children and nonhuman animals distinguish between causal relationships and noncausal correlational relationships. It is also uncontroversial that adult humans in preliterate, scientifically unsophisticated societies share many of our causal beliefs, have a notion of causation that seems much like ours, and engage in counterfactual reasoning. Needless to say, it is unlikely that such people (still less, small children and animals) consciously employ Lewis's similarity metric in making causal judgments or in reasoning with counterfactuals and connecting them to causal judgments. Nor, of course, do

most people in our own society. If people nonetheless make counterfactual judgments in a way that conforms to (or approximately conforms to) Lewis's similarity metric, we face the question of why they do this, what the point or goal behind this practice is. Without an answer to this question, Lewis's theory is "impractical" in the sense described in section 2.1.

One of the advantages of a manipulability theory is that it provides a natural answer to this question. From the perspective of a manipulability theory, the reason Lewis's similarity metric works as well as it does is that the judgments produced by the metric about the truth values of counterfactuals track the judgments produced by the manipulability theory fairly closely. We employ (something like) the similarity metric S1–S4 rather than some alternative because S1–S4 picks out roughly those relationships that are exploitable for purposes of manipulation and control.

I turn now to another respect in which Lewis's theory differs from the manipulationist account. Lewis's similarity criteria are relatively imprecise, at least in comparison with those employed by the manipulability theory. This would be unobjectionable if the truth conditions for counterfactuals (and hence causal claims) were imprecise in the same way, but it is far from obvious that this is so. The passage quoted above in which the criteria S1–4 are introduced naturally suggests a lexicographic ordering: first satisfy S1 as completely as possible, then turn to S2, and so on. However, this picture is complicated by the fact that the features cited in the criteria are vague and their satisfaction admits of degrees. Consider that the exact match of particular fact referred to in S2 can extend over a greater or smaller spatiotemporal region. Does any gain in perfect match, no matter how minimal (perfect match for an additional nanosecond or over an additional  $10^{-10}$  meters) count for more than any localized violation of law? Moreover, how do we count miracles for the purpose of deciding whether we have a single small miracle or a big miracle? Lewis says that big miracles are made up of "many" small ones, but how many is "many"? Are two small miracles enough to make up a big one, or are more required? If just two small miracles are enough for a big miracle, it follows that avoiding them is more important than preserving any region of perfect match of particular fact, no matter how extensive, but that this in turn takes priority over avoiding single miracles. This fundamental difference between one and two miracles is odd enough in itself, but also makes it essential that we have some way of telling whether we have been presented with the former rather than the latter. Moreover, how do we evaluate the extent to which miracles are "alike" or "diverse"? What is the appropriate level or grain to employ in determining whether we have one miracle or many? If there were a single fundamental law that governed all of nature, would all miracles be "alike"? If the gravitational force exerted by the sun switched off for a second, would this be a single, localized miracle or many diverse miracles, as the orbits of many planetary bodies would be affected and the one second of suspension is divisible into many smaller temporal parts? When I consider, under the assumption of determinism, a counterfactual situation in which the pressure of a sample of gas is different from its actual

value, does this require just a single miracle (a single violation of the ideal gas law  $PV = NRT$ ), or does it instead require a huge number of miracles (the kinetic energies of a very large number of individual gas molecules must be changed)?

I do not mean to claim that the criteria S1–S4 are so unclear that they have no definite implications at all; this suggestion is refuted by many of the examples discussed in this section. My point is merely that there are a number of cases in which it is somewhat indeterminate which judgments about the truth or falsity of counterfactuals S1–S4 require. This helps to ensure that S1–S4 often can be construed in such a way that they allow for the reconstruction of our intuitive counterfactual judgments, but skeptics may wonder about the extent to which this is because those judgments are independently dictated (or explained) by S1–S4 and the extent to which this shows only that S1–S4 are sufficiently interpretively flexible (or contain enough free parameters) that usually there will be some way of understanding them that renders them consistent with such judgments.

A final general respect in which Lewis's account differs from my own has to do with the assumption of transitivity. Lewis assumes that causation is transitive when he defines causation as the ancestral of counterfactual dependence; he also requires transitivity to deal with garden-variety cases of preemption, such as the desert traveler case in chapter 2, section 7. By contrast, as explained in chapter 2, I think that it is a mistake to assume that either type or token causation is transitive.

I turn now to some examples in which the judgments about the truth values of counterfactuals delivered by Lewis's similarity criteria and the judgments delivered by the interventionist theory diverge and the latter, not the former, seem intuitively correct. Consider the causal structure in figure 3.6.1.  $C_1$  is a direct cause of each of  $C_2 \dots C_n$ . In addition, there is a direct causal link from  $C_1$  to  $E$ . There are no other causal links. Assume that the occurrence of each cause is sufficient for its effect to occur.

Now consider the counterfactual

(3.6.6) If  $C_2$  and  $C_3$  and  $\dots C_n$  had not occurred, then  $E$  would not have occurred.

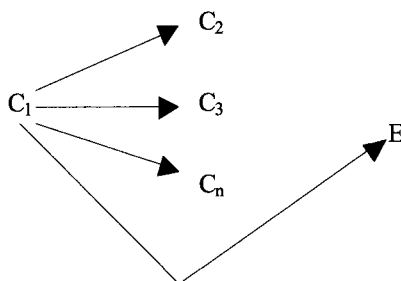


Figure 3.6.1

On the usual understanding of the connection between causation and counterfactuals, (3.6.6) is false because  $C_1$  is a deterministic cause of  $E$ , and even if  $C_2 \dots C_n$  had not occurred,  $C_1$  would have occurred and would have caused  $E$ . On the Lewisian nonbacktracking interpretation of (3.6.6), we consider a possible world that diverges from the actual world in that  $C_2 \dots C_n$  do not occur. This requires the insertion of a miracle someplace. One possibility (world 3) would be that  $C_1$  occurs, but that then  $n-1$  miracles occur, one for each of  $C_2$  through  $C_n$  in such a way that each of  $C_2 \dots C_n$  does not occur. In other words, each of the links between  $C_1$  and each of  $C_2 \dots C_n$  is broken. However, this requires many distinct miracles (a “big” miracle), and S1 tells us to give the greatest weight to avoiding this. (We may suppose that  $n$  is large and that the links from  $C_1$  to each of  $C_2 \dots C_n$  are different from each other.) An alternative possibility (world 4) is to insert a single miracle just before the occurrence of  $C_1$ , so that  $C_1$  does not occur, and in this way the nonoccurrence of  $C_2 \dots C_n$  is ensured, as the antecedent of (3.6.6) requires. If we do this, we have a somewhat less extensive region of match with the actual world than we would under world 3, because  $C_1$  occurs in world 3 and not in world 4, but we only have to introduce one miracle as opposed to the many required in world 3. Because Lewis gives a higher priority to avoiding postulating a number of different miracles than to maximizing match, he seems committed to viewing world 4 as the more similar world to the actual world. However, if world 4 is closest to the actual world, (3.6.6) is true and not false as it should be. Moreover, if world 4 is closest, the following counterfactual is apparently true:

(3.6.7) If  $C_2 \dots C_n$  had not occurred,  $C_1$  would not have occurred.

Hence, because in Lewis's theory, counterfactual dependence is sufficient for causation, it follows that  $C_2 \dots C_n$  cause  $C_1$ . Intuitively, the correct world to look at in evaluating (3.6.6) is not world 4 but world 3, in which (3.6.6) comes out false, despite the fact that world 3 requires many more miracles than world 4.

By contrast, the manipulationist account correctly tells us that in evaluating the counterfactual (3.6.6), we should consider world 3 rather than world 4. As explained above, **IN** commits us to an “arrow-breaking” conception of interventions, according to which, in considering a counterfactual like (3.6.6), we should imagine that all arrows directed into  $C_2 \dots C_n$  are broken. All other arrows, including the arrow from  $C_1$  to  $E$  and any arrows directed into  $C_1$ , are preserved. A process that made the antecedent of (3.6.6) true by removing  $C_1$  would not be an intervention on  $C_2 \dots C_n$  because it would violate the requirement that an intervention must not change other causes of the putative effect variable  $E$  except those causes that are causally between  $C_2 \dots C_n$  and  $E$ . ( $C_1$  is a cause of  $E$  that is not causally between  $C_2 \dots C_n$  and  $E$ .) On the manipulationist account, both the counterfactuals (3.6.6) and (3.6.7) are false, as they should be.

As these last remarks should make clear, the reason the manipulationist account inserts the needed miracle or intervention in the right place is that the characterization of interventions is framed in causal language. This allows us to take account of the relationship between the intervention and the details of the causal structure of the system in which the intervention occurs and to

ensure that the intervention changes only the putative cause variables and those variables, if any, that are caused by these, and that only causal links directed into the variable intervened on are broken. The price of doing this is that we lose the possibility of a reduction, but as the example under consideration illustrates, the attempt to get by with a noncausal theory leads to the insertion of miracles at the wrong points.

This example can be strengthened by making the causal link into  $C_1$  from its cause stochastic, while the links into  $C_2 \dots C_n$  remain deterministic. Then a world in which  $C_1$  (and hence  $C_2 \dots C_n$ ) do not occur need not contain any miracles at all, while the world in which  $C_1$  occurs and  $C_2 \dots C_n$  fail to occur must contain  $n-1$  miracles. In this case, it is even clearer that by Lewis's criteria, the first world is closer to the actual world than the second, even though the second world is the appropriate one to consider in evaluating (3.6.6).

Consider by way of contrast the causal structure in figure 3.6.2, in which there is again a direct causal link from  $C_1$  to  $E$  but in which  $C_2 \dots C_n$  each have independent causes. In this case, (S1–S4) presumably tell us that in evaluating the counterfactual (3.6.6), we should insert miracles that break the causal links (i.e., the arrows) into each of  $C_2 \dots C_n$  from their causes, as there is no alternative way of realizing the antecedent of (3.6.6) that better respects (S1–4). This is also what the manipulationist account tells us to do and both procedures yield the same evaluation of (3.6.6): when evaluated with respect to figure 3.6.2, it is false.

This illustrates an important general difference between Lewis's scheme and the manipulationist picture. On the manipulationist account, when we consider a counterfactual like (3.6.6), we are automatically required to break the causal arrows into each of  $C_2 \dots C_n$  (i.e., to insert a miracle in between each of  $C_2 \dots C_n$  and its causes), regardless of the larger causal structure in which  $C_2 \dots C_n$  are embedded. "Late" miracles, even numerous, are automatically preferred to "early" miracles, even if single. By contrast, in Lewis's theory, whether we break all of the causal links directed into each of  $C_2 \dots C_n$  (i.e., insert many late miracles) or whether instead we break some other link such as the link directed into  $C_1$  in figure 3.6.1 depends on whether  $C_2 \dots C_n$  have many causes or just one. This sort of sensitivity leads to the insertion of miracles in what, intuitively, is the wrong place.

Can Lewis avoid this sort of objection by contending that we do not need to evaluate counterfactuals with complex antecedents like (3.6.6), that all that is

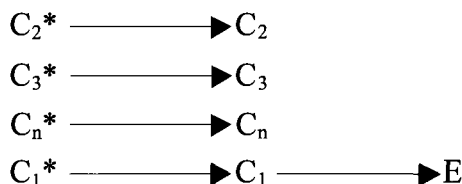


Figure 3.6.2

required is to be able to evaluate counterfactuals with simple antecedents like (3.6.1) and (3.6.2)? This seems unsatisfactory for at least two reasons. First, it is part of our general understanding of the connection between causation and counterfactuals that causal structures often have implications for the evaluation of counterfactuals with complex antecedents involving multiple interventions; for example, we think that (3.6.6) is false when evaluated with respect to the structure in figure 3.6.1. An adequate counterfactual account of causation should capture these implications. Second, as we saw in chapter 2, we need counterfactuals with complex antecedents like (3.6.6) if we are to capture such causal notions as the notion of a direct cause: to say that  $C_1$  is a direct cause of  $E$  in figure 3.6.1 is to talk about what would happen to  $E$  if  $C_2 \dots C_n$  had not occurred and  $C_1$  were (or were not) to occur.

Consider another example.<sup>19</sup> You are driving on an unfamiliar freeway in the left-hand lane when, unexpectedly, the exit you need to take appears on the right. You are unable to get over in time to exit and as a result are late for your appointment. There are only two lanes, left and right. Driving in the left-hand lane (rather than the right) caused you to be late. We think it is true that

(3.6.8) If you had not been driving in the left lane (i.e., if you had been in the right lane), you would not have been late.

On Lewis's theory, when we evaluate (3.6.8), we should consider worlds in which, shortly before you get to the exit, you are in the right-hand and not in the left-hand lane. One such world (world 5) exactly matches the actual world until very, very shortly before you arrive at the exit: you remain in the left-hand lane until immediately before the exit, at which point a miracle occurs and your car dematerializes and reappears instantaneously in the right-hand lane just before the exit. This maximizes match, but there is an obvious problem: it is not at all unlikely that the very occurrence of this miracle will produce effects that will interfere with your exiting. For example, other drivers will be startled and distracted by the sudden appearance of a car in the right-hand lane, perhaps very close to or in contact with cars that already occupy the right lane. Perhaps a collision will occur or other drivers may swerve or slow down, with the result that your exit is impeded. There will also be a great rush of air into the space in the left lane previously occupied by your car, a similar rush as air is displaced from the right-hand lane, and accompanying loud noises, all of which may also interfere with your exit. So, if this is the relevant world to consider, (3.6.8) may well be false, contrary to the result that we want.

We can get rid of all of these additional effects that may interfere with your exiting by postulating additional miracles: the other drivers don't notice when your car suddenly materializes in the right-hand lane, the materialization is coordinated in such a way that there is no collision with other cars, there is no accompanying movement of air or loud noises, and so on. (Call this world 6.) However, this conflicts with requirement S1, which assigns the greatest weight to avoiding large numbers of miracles.

To avoid the untoward results of waiting until the last possible moment to introduce a miracle (or miracles), the obvious alternative strategy is to

introduce it (them) earlier: at some earlier moment, say a quarter mile before the exit, a small miracle occurs in your brain, with the result that you gently and smoothly move your car over to the right-hand lane via normal physical processes that don't involve dematerialization. (Call this world 7.) Lewis ([1979] 1986c, 1986d) seems to explicitly endorse this strategy of avoiding abrupt, discontinuous transitions. He considers the suggestion that, in evaluating the counterfactual "If A were the case, C would be the case," the closest possible world is one that is identical to the actual world at all times before the time at which A occurs and in which A then occurs. He objects that this "makes for abrupt discontinuities. Right up to *t*, the match was stationary and a foot away from the striking surface. If it had been struck at *t* would it have traveled a foot in no time at all? No; we should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future" (1986d, pp. 39–40).

In the example under discussion, world 7 is preferable to world 5 from the point of view of avoiding abrupt discontinuities and providing for an orderly transition to the counterfactual present. However, there are serious problems with the suggestion that world 7 is the appropriate world for evaluating (3.6.8). First, as the quotation from Lewis recognizes, it has the result that the immediate past—the period between the small miracle in your brain and your exit—is counterfactually dependent on and hence caused by your exiting. So, we get backward causation in a case in which backward causation is clearly not at work. Second, and more fundamentally, it is not clear why Lewis is entitled to this strategy. After all, the possible world 5, in which the car remains in the left lane until it materializes at the last possible moment in the right lane, is the one that maximizes actual match. If the materialization of the car in the right-hand lane involves just one miracle, then both world 5 and world 7 contain a single miracle. Thus, Lewis's criteria tell us to prefer world 5 to world 7, and in the latter world, it may very well happen that I fail to exit, so that the counterfactual (3.6.8) is false.

Of course, one might respond that when the car disappears from the left-hand lane and reappears in the right, this involves two small miracles, hence a "big" miracle, rather than a single small one, and that because world 7 involves only a single small miracle, it is closer to the actual world than 5. But this again raises the question of how we are to determine whether world 5 contains two miracles or one, or whether your decision to move to the right a quarter mile before the exit requires only a single miracle. Perhaps the right way to count miracles in the brain is at the level of individual neurons, so that each change in an individual neuron constitutes a distinct miracle. If brains work in such a way that for you to move into the right lane many different neurons must fire differently, this movement will require a largish miracle, after all. I take this to simply illustrate the indeterminacies involved in counting miracles. As remarked earlier, this indeterminacy provides lots of wiggle room when difficulties arise, but it is hardly a virtue.

There are other problems as well with the strategy of preferring earlier miracles and orderly transitions to the counterfactual present to later miracles



and discontinuous transitions. One worry is that once transition periods are countenanced at all, there may be a large number of possible transitions, none obviously closer to the actual world than any other. In other words, there may be many possible worlds, differing in the point at which an early miracle is inserted and what that miracle looks like, that are tied for closest. In some worlds of this sort in which the transition period is protracted, events in the transition period may directly affect whether or not I exit, and if so, the counterfactual (3.6.8) will be false. Suppose that in possible world 8, I move into the right-hand 400 meters before the exit, hit a pothole that damages my car, and continue in the right lane until just before the exit, at which point my car stalls from the damage, and I am unable to exit. By contrast, in possible worlds 9 and 10, in which I move to the right 410 or 390 meters from the exit, I avoid the pothole and successfully exit. Should we conclude that 8 is closer to the actual world than 9 and 10 on the grounds that 8 more closely resembles the actual world in that I fail to exit? That 9 and 10 are closer in that I hit no pothole in the actual world? Are 8, 9, and 10 tied, in which case (3.6.8) is false? What do we say if different numbers of neurons (and hence miracles of different sizes) are involved in the decisions made in 8, 9, and 10? It is hard to see how to provide nonarbitrary answers to such questions, and, intuitively, it seems that the evaluation of the counterfactual (3.6.8) should not turn on how we answer them. The fact that transition periods embroil us in such questions provides an additional reason for avoiding them.

Recall that one feature an intervention on *X* for the purpose of determining whether *X* causes *Y* should possess is that the intervention should not directly affect *Y*: the intervention should change *X*, and it should change *Y* if at all only through *X*. Otherwise, any change in *Y* may result just from the intervention and not as a consequence of any causal link between *X* and *Y*. What the above example illustrates is that nothing guarantees that the miracles that are dictated by Lewis's similarity criteria will possess this feature. That is, there is nothing in Lewis's theory that rules out the possibility that the miracle that changes *X* may itself causally affect *Y* independently of its effect on *X*, and if it does, we may end up with a mistaken evaluation of (3.6.8). This is exactly what happens in world 5: any miracle that makes the car suddenly materialize in the right-hand lane will (or may) have independent effects on the possibility of my exiting. By contrast, this possibility is explicitly excluded by requirement (I3) in our characterization of an intervention, which demands that the process that realizes the antecedent of a counterfactual relating *X* to *Y* must not directly affect *Y* independently of *X*.<sup>20</sup> This requirement implies that the correct possible world to consider in evaluating (3.6.8) is neither world 5 nor world 7, but world 6, the world in which there is no transition period and in which all of the independent effects that the intervention that places my car in the right lane would otherwise have on my exiting are canceled or removed. This is so despite the fact that world 6 involves many miracles, in explicit contravention of the ordering established by (S1–4).<sup>21</sup> In other words, in contrast to Lewis, the interventionist account tells us that we should avoid transition periods entirely (because they may introduce factors that affect the effect independently

of the putative cause), even if the cost of doing so is the postulation of interventions whose occurrence requires many diverse miracles.

I conclude with a final remark about the role played by interventions in the assessment of counterfactuals. It is a striking feature of the kinds of counterfactuals that are relevant to causal and explanatory claims that although we require that they be true when their antecedents are realized by interventions, any more detailed specification of the way their antecedents are realized seems inappropriate and unnecessary. To return to Lewis's example, if a match is a foot from any striking surface and we are asked to consider the counterfactual "If the match were struck, then it would have ignited," it seems misguided to ask, as Lewis in effect does, *how* the match came into contact with the surface (whether it moved instantaneously or, as Lewis seems to prefer, more gradually). Our practices regarding counterfactuals are such that we think that we don't have to answer that question. The notion of an intervention makes this feature of counterfactuals intelligible. Given the way the notion of an intervention has been characterized, if  $I$  is an intervention variable for  $X$  with respect to  $Y$ , it makes no difference to the value of  $Y$  how  $X$  comes to have whatever value it is determined to have by  $I$  or how  $I$  itself comes to possess whatever value it has. This is because it is built into the notion of an intervention that any change in the value of  $Y$  will occur only through the change in the value of  $X$ , so that the details of how both  $I$  and  $X$  come to have their values do not matter to the assessment of counterfactuals of the form "If the value of  $X$  is set to  $x$  by an intervention, then the value of  $Y$  would be . . ." Thus, rather than asking how the match came into contact with the surface, we need only imagine that it came into contact with the surface by some process or other that exerts no effect on whether it ignites, independent of its being struck.

### 3.7 Some Additional Consequences of a Manipulability Theory

I claimed above that one of the attractions of a manipulability theory is that it helps to explain many features of causal concepts that may otherwise seem puzzling and provides a principled basis for distinguishing among competing contentions about the features that causal claims must possess. I turn now to some additional illustrations of this idea.

#### 3.7.1 *Contrastive Focus*

I noted in section 2.5 that a manipulability theory provides a natural account of the role of contrastive focus in causal claims. This has been the subject of a great deal of controversy in the philosophical literature. Do all causal claims (or causal explanations) or only some possess a contrastive structure? Is it the case, as some writers (e.g., Worrall 1984) have suggested, that contrastive structure is a feature of ordinary-language causal talk that is completely dropped in more sophisticated scientific settings? Are there, as Wesley Salmon

(1984, p. 110) contends, legitimate causal claims or causal explanations that are not contrastive at all in the sense that they should not be understood as claims about why a certain kind of outcome rather than one or more alternatives occurred? That is, can we causally explain why  $p$  occurred even though we are unable to explain why  $p$  rather than any alternative  $p'$  to  $p$  occurred? More generally, is contrastive focus a relatively superficial aspect of the linguistic formulation or pragmatics of some but not all causal claims, or is it instead a deeper feature that we should expect to find built into the content of all causal claims?

A manipulationist account of causation suggests definite answers to these questions. Any manipulation of a cause will involve a change from one state to some specific alternative, and how, if at all, a putative effect is changed under this manipulation will depend on the alternative state to which the cause is changed. Thus, if causal claims are to convey information about what will happen under hypothetical manipulations, they must convey the information that one or more specific changes in the cause will change the effect (or the probability of the effect). This in turn means that all causal claims must be interpretable as having a contrastive structure, and it also has the implication (or so I claim in chapter 5) that to causally explain an outcome is *always* to explain why it rather than some alternative occurred. In some cases, contrastive structure is explicitly conveyed by means of the “rather than” locution, as in (2.5.3) (“The occurrence of a blow with momentum  $m_i$  rather than no blow at all caused the chestnut to shatter rather than not to shatter at all”), but in other cases it is implicitly conveyed by other means. For example, the equation (2.5.2)  $F = -k_s X$  implicitly conveys information about contrastive structure because it tells us how, within a certain range, any change or difference in the value of  $X$  (e.g.,  $X$ 's having value  $x$  rather than value  $x^*$ ) will cause a corresponding change or difference in the value of  $F$  (e.g.,  $X$ 's having value  $x$  rather than value  $x^*$  causes  $F$  to have the value  $f$  rather than  $f^*$ ). What is distinctive about equation (2.5.2) is not that it doesn't convey contrastive information but that it conveys, in a single formula, information about the effect on  $F$  of *all* changes or contrasts involving the value of  $X$  (within a certain range of such values). Some such information about how changes in one variable are associated with changes in another is central to all causal claims.

### 3.7.2 Causes and Laws

A long tradition in philosophy contends that there must be a law of nature “underlying” or “backing” all true causal claims. As we shall see in subsequent chapters, some defenders of this idea interpret “law” in such a way that virtually any true causal generalization, regardless of how exception-ridden, vague, or lacking in generality it may be, qualifies as a law. On this interpretation, “Smoking causes lung cancer” and “Penicillin cures tuberculosis” qualify as laws of nature. Call this the weak interpretation of the nomological backing thesis. Other writers adopt a much more demanding conception of “law”: only generalizations meeting various traditional criteria for lawfulness,

such as exceptionlessness, and that resemble certain paradigms regarded as laws by the scientific community will qualify as laws. Call this the strong version of the nomological backing thesis. On the strong version, “Smoking causes lung cancer” is not a law of nature, but the principles of the conservation of energy and momentum arguably are.

What, if anything, do the manipulationist ideas defended above have to say about these contentions? As explained above, the claims embodied in **TC** and **M** about what would happen to *Y* under manipulations of *X* should be interpreted as generalizations, the truth of which requires some degree of reproducibility in the response of *Y* to manipulation of *X*, given sufficiently similar background circumstances. In this sense, type-level causal claims always imply that certain associated generalizations are true. If we interpret the weak version of the nomological backing thesis as claiming only that there is some generality built into type-causal claims, **TC** and **M** support this thesis. Moreover, to the extent that we adopt the account **AC\*** of token-causal claims proposed in 2.7, according to which such claims require for their elucidation reference to background type-causal claims, we can also say that token-causal claims imply that some appropriately related causal generalization is true.

What about the strong version of the thesis? I argue in chapter 4 that although there is nothing in a manipulationist account of causation of the sort embodied in **TC**, **M**, and **AC\*** that implies that this version of the thesis is false, it also does not follow from **TC**, **M**, and **AC\*** that the strong thesis must be true. In other words, the manipulationist account gives us a specification of the content of causal claims that does not require that we assume the strong version of the nomological backing thesis; if the strong version of the thesis turns out to be true, this will be an empirical discovery and not a truth that follows just from the content or meaning of type- (or token-) causal claims.

### 3.7.3 *Causation, Spatiotemporal Continuity, and Unanimity*

What is the relationship between spatiotemporal continuity or connectedness and causation? Is it true that if *Cs* cause *Es*, then tokens of *C* must be spatiotemporally contiguous with tokens of *E* or must be connected by a continuous spatiotemporal process? (Something like the latter claim is suggested by Salmon 1984.) A number of writers have observed that it seems to be logically possible for *C* and *E* to be separated by a spatial and/or temporal gap and connected by no intervening spatiotemporally continuous process and yet for it still to be the case that, if *C* were manipulated, this would change *E*. Both **TC** and **M** will judge that in this sort of case, the relation between *C* and *E* is causal; that is, that spatiotemporal continuity is not necessary for causation.

Philip Kitcher (1989) provides a memorable illustration of this possibility. He imagines a figure who vanishes, quoting verse, “I warmed my hands before the fire of life. It sinks and I . . .” and reappears at some later time (and perhaps in a different place) still quoting, “. . . I am ready to depart.” Suppose that this process is repeated, and it is always found that when earlier stages

are appropriately manipulated (i.e., in a way that satisfies the conditions for an intervention), there are corresponding changes in the later stages. (We somehow intervene to induce the figure to say “To be or not to be,” and as he reappears much later, he is saying “. . . that is the question,” and so on for many other well-known bits of poetry.) Then, according to both **TC** and **M**, it will become very plausible to believe that the later stages of this process are causally dependent on the earlier stages. A manipulability theory in the form of **TC** or **M** will conclude from such examples that it does not follow from the fact that a relationship is exploitable for purposes of manipulation that it must satisfy a spatiotemporal locality or continuity constraint and that when spatiotemporal continuity or contiguity are absent but the relationship nonetheless supports manipulation, we may legitimately take it to be causal.

It is worth noting that a similar conclusion seems to be suggested by the many examples of well-known scientific theories that postulate causal relationships that are not associated with spatiotemporally continuous causal processes—theories that countenance “action at a distance.” Thus, as formulated by Newton, Newtonian gravitational theory claims that one gravitational body can act instantaneously on a second body separated from it by a great spatial distance. It is perfectly true that Newton himself regarded this feature as unsatisfactory or at least as indicating an important incompleteness in his theory, but there seems no reason to deny that his theory describes a causal relationship between the two bodies, and this seems to have been the conclusion reached by most physicists a generation or two after Newton. Partisans of the manipulability conception will think that **TC** and **M** capture in a perfectly adequate way what it means to say that the relationship described by the inverse square law is causal, even though there is no spatiotemporally continuous process linking the bodies: as long as it is true that manipulating the mass or position of the first (second) body will change the gravitational force exerted by the second (first), the inverse square law will describe a causal relationship.

What does seem to be true of the relationship between causation and spatiotemporal continuity is this: putting aside some well-known interpretive problems that arise both in quantum mechanics and General Relativity, it follows, according to the van Dam–Wigner theorem, from Lorentz invariance, that if energy and momentum are conserved in some interaction, they are conserved locally.<sup>22</sup> Hence, if a causal interaction involves transfer of energy-momentum in accord with a conservation law, that interaction will be mediated by spatiotemporally continuous processes that propagate at finite velocity. However, although many causal interactions involve energy-momentum transfer from cause to effect, not all do. For example, cases of causation by omission and by double prevention (see section 5.11) do not. Moreover, both Lorentz invariance and the conservation of energy-momentum are clearly empirical truths and not a priori constraints that follow just from the notion of causation.

Finally, we should note that the same conclusion is also suggested by the naturalistic perspective advocated in chapter 2. Obviously, it is very much in

the interest of both animals and humans that they be able to detect causal relationships in circumstances in which it is not known what the intervening process is or even that one exists. For example, if ingestion of some substance causes illness hours or even days later, it is important to be able to detect this, even though one is unable to trace a continuous spatiotemporal process from ingestion to symptoms. In such a case, it is hard to see what the advantage is of having a notion of causation according to which it is built into the content of causal claims that they must satisfy a spatiotemporal continuity constraint. Again, although it is presumably true in this case that there is such an intervening process, the relevance and usefulness of the information that ingestion causes illness does not depend on its being true (or on its being believed to be true) that there is such an intervening process.

These examples illustrate a more general point: one way of thinking of **TC** and **M** is as a *constraint* on many of the conditions philosophers have claimed are necessary for causation: if these conditions are not necessary for manipulability, there is no motivation for insisting that they are necessary for causation. That is, one can ask whether the relationship between *C* and *E* is such that we can manipulate *E* by manipulating *C* even when these various additional conditions are not satisfied. If so, we may appeal to **TC** and **M** to reject the claim that satisfaction of these additional conditions is required for a relationship to qualify as causal.

As another illustration of this idea, recall the contention, made in connection with probabilistic theories of causation, that if *C*s cause *E*s, then *C* must “unanimously” raise the probability of *E*s across all background circumstances *B<sub>i</sub>*; that is, it must be the case that  $P(E/C \cdot B_i) > P(E/\neg C \cdot B_i)$  for all *B<sub>i</sub>*. We noted in chapter 2 that a probabilistic version of **TC** or **M** will instead reach the conclusion that smoking causes lung cancer as long as it is true, as indeed it is, that there are some interventions on or manipulations of whether some people smoke that change (or raise, if the issue is whether smoking is a positive causal factor) the probability of lung cancer, even if there is also a rare genetic condition among humans such that smoking with this condition fails to raise (or even lowers) the probability of lung cancer. Smoking can be a means to producing (and hence a cause of) lung cancer even though it is not true that all interventions on smoking in all background circumstances raise the probability of lung cancer. Like the demand for spatiotemporal continuity, the unanimity condition looks unmotivated once one takes the connection between causation and manipulability seriously.

All of these examples illustrate an additional important respect in which a manipulability theory can be interesting and nontrivial, despite its failure to furnish a “noncircular” analysis of causation. The point is simply that **TC** and **M** are inconsistent with many widely accepted theories of causation; manipulability theories can be illuminating about causation in part by telling us what causation is *not*. Taking the causation-manipulation link seriously challenges many of the conditions standardly imposed on causal relationships in the philosophical literature, even though that link doesn’t yield anything like a reductive analysis of causation.

### 3.8 Conclusion

The account of causation that I have sketched lacks many of the features of manipulability theories found in the philosophical literature. It is nonreductive, non- (or at least not very) subjectivist, and assigns no special constitutive role to human agency in the elucidation of causal claims. It holds that causal claims of the form  $Xs$  cause  $Ys$  can be true even if human beings lack the power to manipulate  $Xs$  and even if manipulation of  $Xs$  is nomologically impossible. Given that this is the case, it is natural to wonder in what sense the resulting theory still qualifies as a manipulability theory. Hasn't the connection between causation and manipulation become so attenuated that it provides little illumination?

My response consists in reminding the reader of several points made earlier. One way in which considerations having to do with human agency and with our practical interests as agents in manipulating the world play a central role in illuminating our notion of causality is that these interests explain or largely explain why we have a notion of causality at all, and why it takes the form or has the features that it does. In particular, it is our interest in manipulation that explains why we have (or provides the underlying motivation for our having) a notion of causality that is distinct from the notion of correlation. If one asks why we single out those relationships between  $X$  and  $Y$  that persist under interventions on  $X$ , where interventions are understood along the rather complicated lines described in **IN** (why should anyone care about *that*, one might ask), the answer is to be found in our practical interest in changing or controlling nature. Human beings often are (and are justified in believing that they are) in situations in which they can perform actions affecting some variable  $X$  meeting the conditions **IN** on interventions set out above, and in which their interest is in knowing what will happen to some other variable  $Y$  under such an intervention.

It is true and important that we also think of causal claims as holding in circumstances in which human beings will not or cannot carry out the relevant manipulations or in which those manipulations may not even be physically possible. But rather than seeing recognition of these facts as leading away from the core idea of a manipulability theory, I think we should see them as following directly from that idea. The reason for this is that, as emphasized above, it is built into the idea of a relationship between  $X$  and  $Y$  that can be used for purposes of manipulation (i.e., a causal relationship between  $X$  and  $Y$ ) that whether that relationship holds does not depend on whether anyone carries out the manipulation in question or even on whether it is technologically or physically possible to carry out that manipulation. Put slightly differently, it is built into the logic of such a relationship that whether or not it is true that  $Y$  would change under manipulation of  $X$  does not depend on whether  $X$  is in fact manipulated; manipulating  $X$  is not what makes this counterfactual true. The extension of causal talk to circumstances in which manipulation is impossible—long thought to be the fundamental problem facing the manipulability approach—is in fact *required* by the way we think about causation when manipulation is possible.

A second general point is that thinking of causal relationships as relationships that are potentially exploitable for purposes of manipulation enables us to understand why casual claims have many of the features they do and helps to adjudicate between rival claims about those features. For example, it helps to elucidate the differences among different causal concepts, such as the concepts of total and contributing cause. It explains why causal claims involving causes that are unmanipulable for conceptual reasons are typically unclear, and why we may clarify the meaning of ambiguous causal claims by distinguishing among the various hypothetical experiments that may be associated with them. It explains the role of contrast in causal claims. It also clarifies the connection between causation and counterfactuals and explains why Lewis's counterfactual theory works as well as it does. Finally, it illuminates the role of spatiotemporal contiguity in causation and the relationship between causal claims and laws of nature.



## Causal Explanation: Background and Criticism

My discussion in previous chapters has focused on the notion of causation. The focus of the next two chapters is on causal explanation.<sup>1</sup> Contemporary philosophical discussion of this topic begins with the development of a deductive-nomological (*DN*) model by a number of writers, most prominently Carl Hempel in an extremely influential series of papers (1942/1965, 1965a; Hempel and Oppenheim [1948] 1965). These papers and the reaction to them have structured subsequent discussion to an extraordinary degree, and my discussion follows this pattern: I begin with a brief sketch of the model itself and then turn to some well-known objections. My justification for revisiting this very familiar territory is several-fold. First, I think that the real strengths and limitations of the *DN* model and the lessons to be extracted from the problems it faces are still not adequately understood. Second, an examination of these problems will help to prepare the way for the positive account of explanation I propose. This is the subject of the chapter 5.

### 4.1 The Deductive-Nomological Model

I assume that the basic idea of the *DN* model is sufficiently well-known that a detailed description is unnecessary. According to Hempel and Oppenheim ([1948] 1965), an explanation consists of two major “constituents”: an *explanandum*, a sentence “describing the phenomenon to be explained,” and an *explanans*, “the class of those sentences which are adduced to account for the phenomenon” (p. 247). At the most general level, the conditions of adequacy laid down by Hempel and Oppenheim are meant to apply both to the explanation of “general regularities” or “laws,” such as (to use their example) why light conforms to the law of refraction and to particular events, conceived as occurring at a particular time and place, such as the bent appearance of the partially submerged oars of a rowboat on a particular occasion of viewing. Hempel and Oppenheim’s basic idea is that an explanation must take the form of a sound deductive argument in which a law of nature occurs as an essential premise; in their words, “The explanandum must be a logical consequence of the explanans,” the explanans must contain laws that are “actually required for the derivation of the explanandum,” and “the sentences constituting the explanans must be true” (p. 248).

The *DN* model is meant to capture explanation via deduction from deterministic laws. This raises the obvious question of the explanatory status of statistical laws. Do such laws explain at all, and if so, what do they explain, and under what conditions? Hempel (1965a) distinguishes two varieties of statistical explanation. The first of these, *deductive-statistical (DS)* explanations, involves the deduction of “a narrower statistical uniformity” from a more general set of premises, at least one of which involves a more general statistical law. Because *DS* explanation involves deduction of the explanandum from a law, it conforms to the same general pattern as the *DN* explanation of regularities. However, in addition to *DS* explanation, Hempel also recognizes a distinctive sort of statistical explanation, which he calls inductive-statistical or *IS* explanation, involving the subsumption of individual events (e.g., the recovery of a particular person from streptococcus infection) under (what he regards as) statistical laws (such as a law specifying the probability of recovery, given that penicillin has been taken).

Whereas the explanandum of a *DN* or *DS* explanation can be deduced from the explanans, one cannot deduce that some particular individual, John Jones, has recovered from the above statistical law and the information that he has taken penicillin. At most, what can be deduced from this information is that recovery is more or less probable. In *IS* explanation, the relation between explanans and explanandum is, in Hempel’s words, “inductive” rather than deductive, hence the name inductive-statistical explanation. The details of Hempel’s account are complex, but the underlying idea is roughly this: an *IS* explanation will be good or successful to the extent that its explanans confers high probability on its explanandum outcome.

Why suppose that all (or even some) explanations have a *DN* or *IS* structure? Of course, a very great deal has been written on this subject, but in what follows I want to focus on two ideas, both of which play a central motivating role in Hempel’s (1965a) discussion. The first connects the information provided by a *DN* argument with a certain conception of what it is to achieve understanding of why something happens; it appeals to an idea about the object or point of giving an explanation. Hempel writes:

A *DN* explanation answers the question “*Why* did the explanandum-phenomenon occur?” by showing that the phenomenon resulted from certain particular circumstances, specified in  $C_1, C_2, \dots, C_k$ , in accordance with the laws  $L_1, L_2, \dots, L_r$ . By pointing this out, the argument shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon *was to be expected*; and it is in this sense that the explanation enables us to *understand why* the phenomenon occurred. (p. 337; italics in original)

One can think of *IS* explanation as involving a natural generalization of this idea. Although an *IS* explanation does not show that the explanandum phenomenon was to be expected with certainty, it does the next best thing: it shows that the explanandum phenomenon was at least to be expected with high probability. It is in virtue of providing this sort of information that an *IS*

explanation conveys understanding. Hempel appeals to this idea when he introduces the following “general condition of adequacy . . . for any rationally acceptable explanation of a particular event”:

Any rationally acceptable answer to the question “Why did event  $X$  occur?” must offer information which shows that  $X$  was to be expected—if not definitely, as in the case of  $DN$  explanation, then at least with reasonable probability. Thus the explanatory information must provide good grounds for believing that  $X$  did in fact occur; otherwise that information would give us no adequate reason for saying “That explains it—that does show why  $X$  occurred.” (1965a, pp. 367–68)

Put slightly differently, Hempel’s central idea is, in Wesley Salmon’s (1989) words, that “the essence of scientific explanation can be described as *nomical expectability*—that is, expectability on the basis of lawful connections” (p. 57). The reason explanations must have a  $DN$  or  $IS$  structure is that explanatory import or understanding is a matter of exhibiting the explanandum-phenomenon as nomically expectable.

## 4.2 Counterexamples to the $DN$ Model

Is this claim about explanatory import correct? I begin my examination of this question by reminding the reader of some well-known problem cases for the  $DN$  model.

Many explanations exhibit directional or asymmetric features, to which the  $DN$  model, as formulated above, is insensitive. From information about the height ( $h$ ) of a flagpole, the angle ( $\emptyset$ ) it makes with the sun, and laws describing the rectilinear propagation of light, one can deduce the length ( $s$ ) of the shadow it casts on the ground; such a derivation is arguably an explanation of ( $s$ ). It is equally true that, from the length of the shadow, these same laws, and  $\emptyset$ , one can deduce the height ( $h$ ) of the flagpole. This second (“backwards”) derivation (call it Ex. 4.2.1) apparently meets all of the criteria for an acceptable  $DN$  argument, but is no explanation of why the flagpole has this height. Similarly, one can derive the length of a simple pendulum from information about its period  $T$ , the acceleration  $g$  due to gravity that it experiences, and the law  $T = 2\pi\sqrt{\ell/g}$ , but such a derivation is again no explanation (Bromberger 1966).

There are other kinds of explanatory irrelevancies besides those associated with the directional features of explanation. Consider a well-known example due to Wesley Salmon (1971, p. 34):

(Ex. 4.2.2)

( $L$ ) All males who take birth control pills regularly fail to get pregnant

( $K$ ) John Jones is a male who has been taking birth control pills regularly

( $E$ ) John Jones fails to get pregnant

Despite satisfying the requirements of the  $DN$  scheme, ( $L$ ) and ( $K$ ) are no explanation of why Jones fails to get pregnant.

One obvious diagnosis of the difficulties posed by these examples focuses on the role of causal considerations in explanation. According to this analysis, to explain an outcome is to cite its causes. The height of the flagpole causes the length of its shadow, and this is the reason we find a derivation running from the former to the latter explanatory. By contrast, the length of the shadow does not cause the height of the flagpole (the former is rather an effect of the latter), and this is why we don't regard a derivation running from *s* to *h* as explanatory. Similarly, taking birth control pills does not cause Jones's failure to get pregnant, and this is why (4.2.2) fails to be an acceptable explanation (see Salmon 1989 for essentially this assessment).

Although I fully agree with this diagnosis, it does not take us very far. It needs to be supplemented with an account that spells out in detail what a cause (or causal information) is and how the provision of such information contributes to the point or goal of explanation. Salmon has provided such an account in the form of his Causal Mechanical (CM) model, which I discuss in chapter 8. For now, however, I want to draw a simple and uncontroversial conclusion from (4.2.1) and (4.2.2): that these examples show fairly conclusively that the *DN* model, as described above, does not state *sufficient* conditions for successful explanation. Explaining an outcome isn't *just* a matter of showing that it is nomically expectable.

There are two possible reactions one might have to this observation. One is that the idea that explanation is a matter of nomic expectability is correct as far as it goes, but that something more is required as well. The *DN* model (or the more general version of this model understood to encompass the *IS* model as well) does state a necessary condition for successful explanation and, moreover, a condition that is a nonredundant part of a set of conditions that are jointly sufficient for explanation. However, some other, independent feature, *X* (which will account for the directional features of explanation and ensure the kind of explanatory relevance that is apparently missing in the birth control example) must be added to the *DN* model to achieve a successful account of explanation. The idea is thus that Nomic Expectability + *X* = Explanation. Something like this idea is endorsed, for example, by the unificationist models of explanation developed by M. Friedman (1974) and Kitcher (1989), which are discussed in chapter 8.

A second, more radical possible conclusion is that the Hempelian account of the goal or rationale of explanation is mistaken in some much more fundamental way: explanation is not even in part a matter of showing that an explanandum is nomically expectable, and the *DN/IS* models do not even state necessary conditions for successful explanation. That this second reaction is closer to the truth is suggested by two examples due to Michael Scriven (1959a). Both turn on the apparent inability of the *DN/IS* model to accommodate at least some singular-causal explanations.

Suppose that some particular patient, Jones, develops paresis (Scriven 1959a). As it turns out, paresis is caused only by untreated latent syphilis, but only a relatively small percentage of patients with untreated syphilis (roughly 25 percent) develop paresis. As Scriven notes, there seems to be a straightforward

and unproblematic notion of explanation according to which we can explain Jones's paresis by saying that

(4.2.3) Jones's paresis was caused by his untreated latent syphilis.

However, this explanation does not possess an explicit *DN* or *IS* structure. There is no explicit reference to a law, and the condition cited as explaining the paresis not only fails to be nomologically sufficient for paresis, but also fails even to make paresis highly probable. Here we seem to have a case in which an explanandum is explained but by not being shown to be nomically expectable; that is, the case seems to show that nomic expectability is not necessary for explanation.

A second example, also due to Scriven (1959b), makes a similar point:

As you reach for the dictionary, your knee catches the edge of the table and thus turns over the ink-bottle, the contents of which proceed to run over the table's edge and ruin the carpet. If you are subsequently asked to explain how the carpet was damaged you have a complete explanation. You did it, by knocking over the ink. The certainty of this explanation is primeval. It has nothing to do with your knowledge of the relevant laws of physics; a cave-man could supply the same account and be quite as certain of it. Now it is quite true that the *truth* of this explanation is empirical and in this sense it depends on the laws of nature. But its *certainty* has nothing to do with your ability to quote the laws. You have some knowledge about what happens when you knock things over, but so does the cave-man, and this kind of knowledge is totally unlike knowledge of the laws of nature: If you were asked to produce the role-justifying grounds for your explanation, what could you do? *You could not produce any true universal hypothesis* in which the antecedent was identifiably present (i.e., which avoids such terms as "knock hard enough"), and the consequent is the effect to be explained... the explanation has become not one whit more certain since the laws of electricity and inertia were discovered... The simple fact must be faced that certain evidence is adequate to guarantee certain explanations without the benefit of deduction from laws. (p. 456)

Scriven's claim is that

(4.2.4) The impact of *S*'s knee on the desk caused the tipping over of the ink well seems to explain the tipping over of the ink well,

even though (4.2.4) lacks an explicit *DN* or *IS* structure.

Related observations hold for cases involving nonsingular explananda. Recall the example discussed in chapter 3 in which a randomly selected treatment group is exposed to a drug that is withheld from a control group. In this case, it is natural to think that

(4.2.5) The greater expected incidence of recovery in the treatment group is explained by their exposure to the drug.

Even if we agree, for the sake of argument, that this explanation should (somehow) be understood as deductive in structure, it does not appeal, at least explicitly, to any generalization that looks like a law of nature. Another illustration is provided by an example discussed in more detail in chapter 5:

(4.2.6) The qualitative pattern of vote totals  $V$  in New Hampshire elections is explained by the pattern of bias  $S$  in editorial endorsements by the leading New Hampshire newspaper via a regression equation linking  $V$  to  $S$ .

Few would suppose that this regression equation describes a law of nature.

One possible response to these examples is to deny that any of them are explanatory. The problem with this response is that it seems utterly arbitrary and unmotivated. As Hempel acknowledges, people do regularly take singular-causal claims like (4.2.3) and (4.2.4) to be explanatory. Nor do apparent explanations that lack an explicit *DN/IS* structure occur only in “ordinary life”; similar examples can be found in “scientific” contexts as well, as claims like (4.2.5) and (4.2.6) show. Indeed, if we look at the matter from the naturalistic perspective advocated in chapters 1 and 2, it seems very likely that the sorts of explanations emphasized by *DN* theorists—explanations that take the form of explicit deduction from laws—are relative latecomers in human explanatory practice. Both the identification of laws of nature and the exhibition of chains of deductive reasoning require a great deal of intellectual sophistication. Throughout most of human history and probably even today, most explanations look more like (4.2.3) and (4.2.4) than like anything possessing an explicit *DN/IS* structure. Rather than simply dismissing such explanations, a plausible theory of explanation should tell us how they work and what it is about them that provides understanding.

I take this to be Hempel’s view as well. Although he has a number of different and perhaps not entirely consistent things to say about explanations like (4.2.3) and (4.2.4), the most plausible reconstruction of his position is roughly as follows (cf. Hempel 1965a). When one uses a singular-causal explanation of form  $s$  caused  $e$ , one is in some way “tacitly” or “implicitly” appealing to or “presupposing” or “claiming by implication” the existence of some unknown conditions  $K_1 \dots K_n$  and some unknown law  $L$  (linking  $S \cdot K_1 \dots K_n$  to  $E$ , where  $S$  and  $E$  are event-types to which  $s$  and  $e$  belong) such that  $S \cdot K$  and  $L$  are premises in a *DN* or *IS* explanation of  $E$ . Thus, in the case of (4.2.4), there will be an “underlying,” or “presupposed” *DN* explanation<sup>2</sup> having something like the following structure:

(L) Whenever  $S \cdot K_1 \dots K_n$ , then  $T$   
 (4.2.7) (C)  $S \cdot K_1 \dots K_n$   
 (E)  $\therefore T$

where  $S$  = a knee impacts the desk,  $T$  = the inkwell tips over.

Of course, (4.2.4) does not by itself tell us what these additional conditions  $K_1 \dots K_n$  and the unknown law  $L$  are. Nonetheless, we may think of (4.2.4) as

conveying, albeit in an incomplete or imperfect form, information about what the ideal underlying explanation (4.2.7) looks like. In particular, we may think of (4.2.4) as conveying part of the information conveyed by (4.2.7)—“part” in the quite literal sense that the cause event cited in (4.2.4) is one of the conjuncts in the antecedent of the law  $L$  in (4.2.7) and the effect event cited in (4.2.4) is the explanandum of (4.2.7). Hempel uses various terms to describe this idea; he speaks of (4.2.4) and other similar explanations as “partial” or “elliptical” explanations or as “explanation sketches,” but a common overall strategy runs through his discussion. As Philip Kitcher (1989) puts it, this is “to distinguish between what is said on an occasion in which explanatory information is given and the ideal underlying explanation” (p. 414). The former (e.g., (4.2.4)) somehow points to, invokes, or conveys information about the latter (4.2.7), and it is because this is so (or to the extent that this is so) that the former counts as explanatory. Presumably, Hempel would advocate a similar strategy in connection with (4.2.5) and (4.2.6).

But why think that an explanation of form (4.2.7) “underlies” or is “presupposed by” (4.2.4) (the impact of the knee caused the tipping of the inkwell) to begin with? The main argument for this claim, both in Hempel’s work and elsewhere in the philosophical literature, appeals to broadly Humean considerations regarding the nature of causation. According to this argument, claims like (4.2.4) that explicitly contain the word “cause” or cognate causal locutions (“produce,” “bring about,” “resulted from,” etc.) stand in need of analysis or elucidation: they cannot be left as primitives, but instead must be translated into claims that do not themselves employ explicit causal locutions. Following Hume, Hempel supposes that an illuminating, noncircular analysis of causal claims will appeal to facts about the holding of regularities of some sort or (what Hempel takes to be the same thing) “laws.” This is why a singular-causal claim like (4.2.4) “implies” or “presupposes” the claim that some associated regularity or law holds. Hempel then assumes that it follows that we can then think of this regularity as contributing to the explanation provided by the singular-causal claim.

We find this line of thought, which constitutes the second of the two major arguments to which Hempel appeals to motivate the *DN* model (the first being the connection between understanding and nomic expectability), in Hempel’s (1965a) comments on Scriven’s inkwell example (4.2.4):

Undeniably, in our everyday pursuits and also in scientific discussions, we often offer or accept explanatory accounts of the sort illustrated by Scriven’s example. But an analytic study of explanation cannot content itself with simply registering this fact: it must treat it as material for analysis; it must seek to clarify what is *claimed* by an explanatory statement of this sort, and how the claim might be *supported*. And, at least to the first question, Scriven offers no explicit answer. He does not tell us just what, on his construal, is asserted by the given law-free explanation; and it remains unclear, therefore, precisely what claims he regards as having primeval certainty, for cave-man and modern physicist alike. Presumably the explanation he has in mind would be expressed by

a statement roughly to the effect that the carpet was stained with ink because the table was knocked. But, surely, this statement claims by implication that the antecedent circumstances invoked were of a kind which generally yields effects of the sort to be explained. Indeed, it is just this implicit claim of covering uniform connections which distinguishes the causal attribution here made from a mere sequential narrative to the effect that first the table was knocked, then the bottle tipped over, and finally the ink dripped on the rug. Now, in a case such as the spilling of the ink, we feel familiar, at least in a general manner, with the relevant uniform connections even though we may not be able to state them precisely, and thus we are willing to take them for granted without explicit mention. (pp. 360–61)

Hempel's contention that singular-causal explanations like the paresis example (4.2.3) and the inkwell example (4.2.4) are implicit *DN* or *IS* explanations or sketches of such explanations might seem at best relevant to the question of whether the *DN* model provides an adequate reconstruction of this particular sort of explanation. In fact, however, how we assess this contention has far more general implications. Many of the basic elements of Hempel's treatment are reproduced in more recent discussions, often in contexts that have little to do with singular-causal explanation. In particular, Hempel's overall strategy of trying to understand how explanations like (4.2.3) and (4.2.4) work by treating them as devices for conveying information, but in a "partial" or "incomplete" way, about underlying "ideal" explanations of a *prima facie* quite different form that are at least partly epistemically hidden from those who use the original, nonideal explanation has continued to be very popular in recent theorizing about explanation. This strategy, which I will call the "hidden structure" strategy, forms the basis, for example, of Peter Railton's (1978, 1981) "ideal text" approach to explanation (on which more below) and also plays an important role in the accounts of explanation developed by Philip Kitcher (1989) and David Lewis (1986a). It also forms the basis of a great deal of contemporary discussion of how so-called *ceteris paribus* generalizations in the special sciences function in explanation, a standard idea being that such generalizations, although themselves full of exceptions, somehow explain in virtue of conveying information about underlying laws, unknown to most or all users, which *are* exceptionless (see chapter 6).

Many philosophers will regard some version of the hidden structure strategy as the obvious and natural way to understand how explanations like (4.2.3)–(4.2.6) work. My contrary view is that the hidden structure strategy (or at least the way that strategy is used both by Hempel and by more recent theorists) is highly problematic. Because this claim plays a role in motivating the ideas developed in subsequent chapters, it will be worthwhile to consider in some detail the problems the strategy faces.

Assessing the hidden structure strategy is important for a number of reasons. First, the strategy seems to be the only plausible way of protecting the *DN/IS* model against counterexamples of the sort represented by (4.2.3)–(4.2.6). Moreover, if the hidden structure strategy fails, the difficulties that



explanations like (4.2.3)–(4.2.6) present for the *DN* (and *IS*) models will equally be difficulties for more recent models of explanation as well. As an illustration, consider again the unificationist account of explanation advocated by Michael Friedman and Philip Kitcher. Although Kitcher, at least, is willing to countenance explanations that appeal to general unifying principles that fall short of being laws, it is fair to say, as a first approximation, that the unificationist approach retains the basic commitments of the *DN* model and simply adds to them. Thus, according to the unificationist approach, explanation always involves a deduction or derivation from a generalization of considerable scope or generality. Roughly speaking, Kitcher's strategy is to attempt to deal with counterexamples to the *DN* model such as the flagpole example (4.2.1) and the birth control pills example (4.2.2) by adding an additional restriction to the basic *DN* framework; his idea is that if we require that successful explanations unify in the right way, we can avoid these counterexamples. I argue in chapter 8 that this strategy fails in connection with (4.2.1) and (4.2.2). However, even if one accepts Kitcher's treatment of these examples, the strategy of adding an additional condition to the *DN* analysis does not help address the objection that examples like (4.2.3)–(4.2.6) show that possession of a *DN* (or *IS*) structure is not a *necessary* condition for successful explanation. To defeat this objection, some version of the hidden structure strategy seems to be required. In particular, on the assumption that the hidden structure strategy fails and that explanations like (4.2.3)–(4.2.6) are genuine explanations just as they stand and not in virtue of what they convey about a hidden structure, it follows that the unificationist model (as a strengthened version of the *DN* model) also fails to state a necessary condition for successful explanation. In other words, if the hidden structure strategy fails, then what (4.2.3)–(4.2.6) show is that explanation need not always be deductive in structure and need not always appeal to laws or other principles of great generality.

A similar point holds for what, again following Peter Railton (1981), we might call *nomothetic* models of explanation. Such models may take a number of different forms, but they share the common idea that laws play a central role in all explanation—the idea is that even if explanation is not (or not always) a matter of exhibiting nomic grounds for expecting, it is nonetheless true that all forms of explanation involve subsumption of explananda under some law or laws, either overtly or by “providing information about” some explanation that does this. Such views are also undermined if we cannot see explanations such as (4.2.3)–(4.2.6) as explaining in virtue of possessing a hidden nomothetic structure.

More generally, the hidden structure strategy raises a number of important issues about the role of epistemic considerations in explanation and the representation of causal relationships that need to be addressed by any theory of explanation. Is it legitimate to think of a theory or set of claims  $T_1$  as explanatory in virtue of conveying information about some other theory or structure  $T_2$ , even though users of  $T_1$  know nothing about  $T_2$  and may not even know that it exists? To what extent does successful explanation require the actual

exhibition of relationships that investigators can recognize as obtaining rather than the mere existence of such relationships independently of anyone's knowledge? Such questions lurk just below the surface in many contemporary discussions of explanation. I explore them in what follows.

### 4.3 Four Theses about the Relationship between Causal Claims and Laws

As a softening-up exercise, let me distinguish, more explicitly than I have hitherto, among several projects that might be associated with a theory of causal explanation. The first has to do with the first-order *scientific* task of constructing a causal explanation of the phenomena in some domain of inquiry. Here, the task is to represent the causal relationships that govern the phenomena in a way that we can comprehend. Suppose that in fact there is some additional dichotomous variable  $K$  such that if one has untreated latent syphilis ( $U$ ) and one value for  $K$ , one always develops paresis ( $P$ ), and if one has untreated latent syphilis and the other value for  $K$ , one never develops paresis; that is, with the inclusion of  $K$ , the system is fully deterministic. Obviously, whether  $K$  exists is completely independent of what ordinary users of (4.2.3) believe about the existence or nonexistence of  $K$ . And, assuming that the values of  $K$  represent legitimate properties that are not intractably disjunctive or complex (see below), if we are just interested in modeling the causes of paresis, it is legitimate to do so by means of a deterministic equation (e.g.,  $U \cdot K = P$ ), regardless of what ordinary users believe.

The scientific project just described should be distinguished from the *interpretive* project of describing or representing the content or meaning of claims like (4.2.3) or of identifying the information that (4.2.3) provides in virtue of which it is taken to be explanatory. I take the disagreement between Scriven and Hempel, and the question of the acceptability of the hidden strategy more generally, to have to do with this interpretive project (and with a related normative issue that I come to below), rather than with the scientific project: Hempel claims and Scriven denies that (4.2.3) conveys information about an implicit *DN* or *IS* structure and is explanatory for this reason. Suppose that most ordinary users of (4.2.3) do not believe that the additional condition  $K$  exists, or that it does not occur to them to even consider whether  $K$  exists, or that they have no evidence one way or another whether  $K$  exists, and that they do not believe that the goodness of the explanation provided by (4.2.3) depends on whether  $K$  exists, and do not take themselves to be committed to the existence of condition  $K$  when they use (4.2.3). Scriven's contention is that if this is the case, then even if condition  $K$  does in fact exist, we are not entitled to conclude, at least without considerable additional argument, that the *DN* or *IS* argument that Hempel would associate with (4.2.3) (which would appeal to  $K$  and some law linking  $U$  and  $K$  to  $P$ ) captures what (4.2.3) says or what people are committed to when they use (4.2.3), or that (4.2.3) is explanatory because it conveys information about this argument. In other words, we cannot conclude,

simply on the grounds that *K* exists, that claims about the existence of *K* must be somehow “implicit” in (4.2.3). If we want to defend this interpretive claim about (4.2.3), some *additional* argument is required. For example, it might be argued that insofar as (4.2.3) has any clear meaning at all, then the only possible interpretation is that it claims that a backing law involving *K* exists. I take this to be one way of reconstructing the argument Hempel presents in the passages quoted above.

Some readers may wonder why it matters, for the purposes of constructing a theory of causal explanation, whether claims like (4.2.3)–(4.2.6) should be interpreted as implicit *DN* or *IS* arguments. Isn’t the only real issue the “scientific” one of whether the additional conditions *K* and the backing law linking *U* and *K* to *P* exist?

In fact, however, there is a *prima facie* plausible argument, which I take Scriven to be making, that this interpretive issue bears directly on the adequacy of the *DN* model. The argument is simply that claims like (4.2.3)–(4.2.6) do seem to explain, and if they do not do this by in some way conveying the information that the *DN/IS* model claims is required for successful explanation (i.e., information about a nomologically sufficient or at least a probabilifying condition), then the *DN/IS* model must be incorrect in claiming that this information is necessary for successful explanation. This, at any event, is the argument that I examine in what follows.

The interpretive issue just described also bears on the *normative* issue of specifying the information that “ought” to be included in successful explanations. Consider the explanations that are actually on offer for the behavior of complex systems of the sort studied in disciplines like biology and economics, for example, explanations of the behavior of second-price auctions or the development of *Drosophila*. Suppose that it is true that, as I suggest below, these explanations do not appeal to anything that looks much like “laws” or nomologically sufficient conditions. If so, and if there are principled reasons why the sort of information required by the *DN* model may be difficult or impossible to obtain in the case of such systems, this will provide at least some motivation for the suggestion that the normative criteria we use to evaluate such explanations should not imply that we have not successfully explained unless we have satisfied the *DN* criteria.

To expand on this point, consider that the complex systems mentioned above are of course physical systems whose behavior is governed by fundamental physical laws. In this sense, it is uncontroversial that there is some nomothetic structure “underlying” their behavior. Nonetheless, it is also uncontroversial that we cannot understand these systems by describing the characteristics of each microphysical constituent of them and then writing down and solving the equations governing all of their interactions. Among other things, we cannot possibly gather all of the information we would need about the characteristics of the microphysical constituents. Moreover, the relevant equations would be completely computationally intractable. If we are to explain the behavior of such systems, we must be able to represent the causal relationships obtaining in them in such a way that investigators can gather evidence relevant to

whether those relationships obtain. We also must be able to represent such relationships in a way that makes it possible to trace out their implications, for example, by solving analytically the equations describing them or by carrying out approximations or simulations. This, of course, is just what biologists and economists aim to do. Whatever may be going on in complex biological and economic systems at the level of the relationships postulated by fundamental physics, it is not through recognition of fundamental physical relationships that we understand how such systems behave. This by itself suggests that there must be some marks or criteria of successful explanation in these disciplines that are specifiable or recognizable independently of facts about fundamental physics and other “underlying” but inaccessible theories. It is at least a legitimate goal for a theory of explanation to specify what these criteria are. And to the extent that underlying DN or nomothetic structures exist only at the level of physics and chemistry, we don’t want an account of explanation that implies that we have achieved nothing, from the point of view of explanation, unless we have exhibited such structures. Again, the hidden structure strategy cannot be validated simply by observing that such structures will always exist; it is an additional, independent question whether successful explanation requires that we exhibit them.

With these remarks as background, I turn now to a more detailed exploration of some of the issues raised by the hidden structure strategy. I begin with the idea, central both to Hempel’s argument and to views of many other philosophers, that associated with all true causal claims (whether singular or general) there will be some “underlying” law or laws. It will be useful to distinguish four different versions of this idea. The first commits us to nothing more than the formulation in the second sentence of this paragraph: for every true causal claim an underlying law exists. Call this the *underlying* thesis. This should be distinguished from a second thesis about the *meaning* (or, in the formulation of some writers [e.g., Davidson 1967], who take the two to be linked) the truth conditions of causal claims. According to the meaning thesis, it is part of the meaning of every causal claim that some associated, “underlying” law must be true; the truth of the causal claim *entails*, in virtue of its meaning, the truth of some underlying law. The meaning thesis thus agrees with the underlying thesis that an associated law will always exist, but adds to this the claim that the existence of the law is somehow built into or entailed by the meaning of the causal claim. It is this thesis that Hempel seems to be endorsing when he says, in the passage quoted above, that Scriven’s causal claim about the tipping over of the inkwell “claims by implication” that a certain uniformity holds. A similar thesis is endorsed by Donald Davidson. He writes in his well-known essay (1967), that “the relation in general [between singular-causal claims of the form ‘*a* caused *b*’ and the laws that ‘back’ them] is this: if ‘*a* caused *b*’ is true then there are descriptions of *a* and *b* such that the result of substituting them for ‘*a*’ and ‘*b*’ in ‘*a* caused *b*’ is entailed by true premises of form (*L*) and (*P*) [see below] and the converse holds if suitable restrictions are put on the descriptions” (p. 700). Here, (*P*) is a premise that asserts the occurrence of the cause, and (*L*) gives the general form of the causal

law that backs singular-causal claims. According to Davidson, this backing law ( $L$ ) will take the form of a conjunction:

$$(S) (e)(n)((Fe \ \& \ t(e) = n) \rightarrow (\exists!f)(Gf \ \& \ t(f) = n + \epsilon \ \& \ C(e, f)))$$

and

$$(N) (e)(n)((Ge \ \& \ t(e) = n + \epsilon) \rightarrow (\exists!f)(Ff \ \& \ t(f) = n \ \& \ C(f, e)))$$

Other philosophers, although skeptical of the meaning thesis, subscribe to some form of the underlying thesis. For example, Peter Railton (1981), as I understand him, is not committed to the meaning thesis, but he clearly accepts a version of the underlying thesis: he thinks that associated with every true causal claim is an “ideal explanatory text” that does appeal to laws.

Both the underlying and meaning theses should be distinguished from the *epistemological* thesis: that it is only by knowing the law or regularity underlying a causal claim that can one reliably tell whether the claim is true. On this view, we require knowledge of laws to reliably distinguish genuinely causal sequences from merely accidental sequences or noncausal correlations.

All three of the previous theses should also be distinguished from the following thesis about (causal) *explanation*: laws are a necessary part of or make an essential contribution to every explanation. For a putative explanation to have explanatory import or to provide understanding, it must either explicitly cite some law or else implicitly convey information about the existence of a law as, for example, on Hempel’s analysis, the claim about the tipping of the inkwell (4.2.4) implicitly conveys information about the existence of the laws in (4.2.7). When Hempel claims that causal explanations like (4.2.3)–(4.2.4) are “implicit” or “partial” *DN* or *IS* explanations, I take him to be committed to this fourth thesis about the role of laws in explanation, and not just to the meaning, underlying, or epistemological theses. It is the explanation thesis, and not just the first three theses, that must be established if Hempel is to make good on his claims about the connection between explanation and nomic expectability.

Which, if any, of these four theses is correct and what are their logical interrelations? I think it is fairly clear that Hempel is committed to all four and that he often writes as though they are equivalent or interchangeable. His general strategy in the passages quoted above is to argue for one or both of the meaning and epistemological theses, and then to regard these as tantamount to the (explanation) thesis, which, as remarked above, is the conclusion he ultimately wishes to establish. Many other philosophers argue similarly.<sup>3</sup> Indeed, many of the common formulations of the connection between causal explanation and the existence of laws that one finds in the philosophical literature—that laws are always “involved in” or “required for” or “presupposed by” causal explanations—are systematically ambiguous among the theses distinguished above. I suspect that many readers who are well aware of the difficulties facing the *DN* model will nonetheless agree with Hempel in holding

that all four theses are correct and will also think that the theses themselves are closely interconnected, perhaps so closely that it is not worthwhile to try to distinguish among them. My own view is that both the epistemological thesis and the explanation thesis are false, and that the meaning thesis is at least dubious. Only the underlying thesis is plausible, and even it is unclear in crucial respects. I also think that the theses are largely logically independent of each other and that the explanation thesis in particular does not follow from either the underlying or epistemological thesis. It follows that one cannot appeal to either thesis to motivate the claim that all causal explanations explain in virtue of tacitly or implicitly invoking or providing information about a law.

I turn first to the underlying and meaning theses. As stated, both theses are unclear in a number of respects. To begin with, putting aside the quotation from Davidson, nothing has been said yet about what it is for a law to be “associated with” or to “back” or “underlie” a causal claim. As we will see, it is far from obvious how to unpack the quoted phrases. However, I take the intended sense to be something like this: for the underlying or meaning thesis to be true, some law must hold that bears an appropriate relationship to the causal claim. For example, it would hardly be satisfactory, from the point of view of the meaning thesis, if the only law entailed by the claim about the tipping of the inkwell (4.2.4) was a law governing stellar development. Rather, the entailed law should be a law that connects impacts of knees (or some redescription of these) to tipping over of inkwells. It should be of the right sort to serve as a “ground” or “basis” or “truth maker” for the causal claim that entails it.

A second and related point is that, as formulated, both the underlying and meaning theses say nothing specific about what the laws or regularities associated with various sorts of causal claims look like. Hempel often writes as though he expects that the laws underlying various singular-causal claims will be stated largely or entirely in the vocabulary of (or, if we may put it this way, at the explanatory “level” of) those claims. Thus, in the inkwell example, his idea is apparently that there will be a law or “uniformity” linking knee movements to the tipping over of inkwells to carpet stains; in the case of the claim that burning of the haystack was caused by the dropping of the cigarette, there will be a law specifying the conditions under which dropping of lighted cigarettes is inevitably followed by fires, and so on. Let us call a law or generalization bearing this sort of relationship to a causal claim a *direct generalization* of that claim.<sup>4</sup>

A number of other commentators who accept the meaning thesis, including, most notably, Davidson (whose views are adumbrated in the passage quoted above), have contended that it is implausible to suppose that the laws associated with causal claims like (4.2.4) will be direct generalizations of these claims. They hold that there are no laws linking knee movements to the capsizing of inkwells or the dropping of cigarettes to fires. Instead, they suggest that the laws underlying many causal claims will be stated in a quite different vocabulary and will hold at a quite different explanatory level from the original

claims themselves. For example, in the case of (4.2.4), the relevant laws might be various laws of Newtonian mechanics (the conservation of kinetic energy and linear momentum) governing the collision between the knee and the table, the law describing the gravitational force incident on the inkwell as it falls from the table, laws describing the chemistry of the ink and how this functions in staining, and so on. When the law or laws underlying a causal claim are stated in a different vocabulary from the claim, I will say that the laws are *indirect generalizations* of the causal claim. It follows, given this view of the character of the underlying laws, that to the extent that it is plausible that an implicit *DN* argument of form (4.2.7) “underlies” (4.2.4), its nomological premises will consist of laws about the conservation of momentum and so on, not laws that relate knee movements to tipplings of inkwells.

Two other consequences of the view that the laws underlying many causal claims are indirect generalizations are also worth noting. First, though it is arguable, even if not obviously correct, that if the laws associated with causal claims are direct generalizations of those claims, those generalizations will usually be known to ordinary users of the causal claims, the corresponding contention is completely implausible if the associated laws are Davidsonian indirect generalizations. That is, it seems undeniable that if the laws underlying (4.2.4) have to do with the conservation of momentum and so on, then these laws are unknown to many who presently employ (4.2.4) and were unknown to *all* users in the not-too-distant past, before these laws were discovered. Thus, if the laws underlying causal claims are typically indirect generalizations, the epistemological thesis looks dubious and certainly does not follow from the underlying or meaning theses. Second, if the only laws underlying (4.2.4) are indirect generalizations, the initial suggestion I made above regarding what it might mean to say that (4.2.4) “conveys information” about the underlying *DN* argument (4.2.7) will no longer work: we can’t say that (4.2.4) makes use of or shares some of the very same descriptions (“the movement of S’s knee,” etc.) that are involved in the underlying *DN* explanation (4.2.7), and that this is why it is correct to think that (4.2.4) conveys information about (4.2.7). If (4.2.4) really works by conveying information about some unknown underlying explanation of form (4.2.7), this must be in virtue of some other notion of “conveying information about.” I return to both points below.

Is it plausible that there are direct generalizations that qualify as laws of nature associated with all or most causal claims? This question draws our attention to another respect in which both the underlying and meaning theses are unclear: there is little consensus about the criteria for lawfulness. As I remarked in chapter 3, some philosophers understand the notion of “law” in such a way that virtually any true causal generalization, regardless of how exception-ridden, vague, or imprecise it may be, qualifies as a law. On this interpretation, “Impacts on containers of liquids cause spills,” “Smoking causes lung cancer,” “Latent syphilis causes paresis,” and “Penicillin cures tuberculosis” are laws of nature. Although I suggest below that there are good reasons not to adopt such a permissive notion of law, I also think that if this is what we decide to mean by

“law,” it is plausible that all causal claims, including singular-causal claims, entail the existence of laws. This, at any event, was the view adopted in chapter 2, where the content of both type- and token-causal claims was understood in terms of claims about the holding of generalizations.

The permissive notion of law just described contrasts with a much stricter notion that is common in both philosophical and scientific usage. A standard view in philosophy of science has been that laws must have many or all of the following features: they are universally quantified conditional statements, describing exceptionless regularities that contain no reference to particular individuals, times, or places, contain only “purely qualitative” predicates, are confirmable by their instances, and support counterfactuals. Presumably, to qualify as a law, a generalization also must relevantly resemble paradigmatic examples that are recognized as laws in scientific practice (e.g., Maxwell’s equations, the field equations of General Relativity, etc.). It is notorious that there is considerable debate about many of the philosophical criteria; it is widely doubted (in my view, with good reason) that they successfully distinguish between laws and accidental generalizations and that they fit paradigmatic examples of generalizations that are described as laws in scientific practice.

I explore some of the issues raised by this debate in chapter 6. There I argue in favor of a notion of law that is restricted in such a way that generalizations that are as exception-ridden and imprecise as

(4.3.1) Smoking causes cancer

do not count as laws. However, even if I am wrong about this, it seems uncontroversial, for the purposes of the present discussion, that the relevant notion of law should be restricted to exclude generalizations like (4.3.1). The reason for this is that generalizations like (4.3.1) cannot, as they stand, play the kind of role in explanation that is assigned to laws by the *DN* model and its modern descendants. In particular, if laws are to serve as nomological premises in deductive arguments, they must at least be exceptionless, and (4.3.1) plainly is not. Similarly, a generalization such as

(4.3.2) Latent syphilis causes paresis

cannot be used as a nomological premise in an *IS* argument designed to show, for example, that particular instances of paresis “were to be expected” because (among other things) that generalization says nothing about the conditions under which the probability of paresis is high. Instead, if the *DN/IS* account or its modern descendants are to be defended, explanations that appeal to generalizations like “Smoking causes cancer” and “Latent syphilis causes paresis” must be reinterpreted as devices for conveying information about generalizations that are suited for playing the role of nomological premises in *DN/IS* explanations, as in the hidden structure strategy.<sup>5</sup> For this reason, I interpret the notion of “law” in the remainder of this chapter in such a way as to exclude exception-ridden generalizations like (4.3.1) and (4.3.2).



#### 4.4 Do Causal Claims Entail Laws?

With this as background, let us return to the question of whether every true causal claim entails a backing direct generalization that is a law. Two initial points seem clear. First, in the case of many causal claims, such as (4.2.3) (“Jones’s syphilis caused his paresis”) and (4.2.4) (“The impact of the knee caused the inkwell to tip over”), it is widely acknowledged that it is at least difficult to formulate genuinely exceptionless direct generalizations. Inkwells do not always spill when the tables on which they are resting receive knocks, and it is far from obvious how to write down a generalization, formulated in the vocabulary of (4.2.4), that excludes all the nonconforming cases. Similarly for (4.2.3), (4.2.5) (treatment with drug explains recovery), and (4.2.6) (editorial slant explains voting patterns in New Hampshire). If only exceptionless generalizations count as laws, then the thesis that for every true causal claim there is an associated direct generalization that qualifies as a law is not self-evidently correct. Second, exception-ridden generalizations like (4.3.1) and (4.3.2) are certainly not described as laws in scientific texts and papers. Thus, when one claims that every causal claim entails some direct generalization that qualifies as a law, one cannot claim to be drawing on an intuitively clear notion of lawfulness that is widely accepted in scientific practice. Again, some argument (other than that if such generalizations *were* laws, this would be a way of saving the DN model) is required to explain why the notion of law should be extended to cover such generalizations.

These considerations do not, of course, conclusively demonstrate that it is mistaken to think that there will be direct generalizations of claims like (4.2.3)–(4.2.6) that qualify as laws, but I do take them to suggest that Davidson’s contention that the laws, if any, that underlie many causal claims are likely to be indirect rather than direct generalizations of those claims is *prima facie* more plausible. It is worth taking this indirect generalization view seriously and exploring its implications for the theses under discussion.

I also noted above that if the underlying and meaning theses are to have a clear sense, an account needs to be provided of what it is for a law to “underlie” or “back” a causal claim. I turn now to a more detailed exploration of this issue. This problem needs to be addressed even by those who think that every causal claim entails some *direct* backing generalization that qualifies as a law, because, among other things, the notion of a generalization employing the same “vocabulary” or being at the same “level” as a causal claim is far from clear. However, the problem is far more acute for those who think that the laws whose existence is entailed by causal claims often will be *indirect* generalizations of those claims, for in this case, one faces the problem of characterizing what the required “backing” relationship is between claims stated in two different vocabularies. As is well-known, Davidson has a very specific proposal, sketched in the passages quoted above, about what the backing relationship involves in the case of singular-causal statements: singular-causal statements relate events, laws of nature are regarded as universally quantified conditional statements with the quantifier ranging over events, and a law “underlies”

a singular-causal claim when the events related by the singular-causal claim “instantiate” (satisfy the antecedent and consequent) of the law.

There is much that is problematic about this proposal. It is not clear that singular-causal claims describe relations between events.<sup>6</sup> The proposal relies on a dubious criterion of event identity.<sup>7</sup> Moreover, if the proposal is to adequately capture the “truth conditions” for singular-causal claims, it requires highly specific assumptions about the “logical form” of the underlying laws, namely, that they specify the conditions under which one and only one event of a certain sort will occur. (See the passage from Davidson quoted above, in which laws are understood as claims that, under the conditions specified in their antecedents, unique events of a certain sort will occur.)<sup>8</sup> There are good reasons to doubt that real laws have anything like this form. Indeed, it is implausible that any purely syntactic characterization—whether framed in terms of “instantiation” of the antecedent and consequent of a universally quantified conditional claim or framed in some other terms—will capture the “underlying” relation, if only because equivalent laws may be given different syntactical formulations. In addition, Davidson’s proposal addresses only singular-causal claims. It does not even attempt to provide truth conditions for garden-variety type-causal claims such as “Smoking causes lung cancer,” which (implausibly) apparently require underlying laws with a different logical form.<sup>9</sup>

There are also many reasons to doubt that the ontology of typical singular-causal claims will mesh with the ontology of the laws that underlie them in anything like the simple way required by Davidson’s proposal. Laws of nature often take the form of differential equations describing how quantitative magnitudes change in a continuous way, with no spatiotemporal gaps. Singular-causal claims often relate variables or properties that take only discrete values and are separated by spatiotemporal gaps. Moreover, even if we agree that both singular-causal claims and laws of nature relate events, why suppose, as Davidson does, that there is a single underlying law or that the underlying scientific framework posits single events to be identified with the cause-and-effect events in the original causal claim? Often, the events related in a singular-causal claim will correspond to what, from the perspective of the underlying scientific framework, are complex, spatially and temporally distributed, gerrymandered and unnatural-looking congeries of events falling under many different laws.<sup>10</sup> This is presumably true, for example, of singular-causal claims like (4.4.1) “Exposure to asbestos caused Jones’s lung cancer,” (4.4.2) “Smith’s desire for water caused him to drink,” or (4.4.3) “The general price inflation of the early 1970s in the United States was caused by the OPEC oil embargo.”<sup>11</sup> “Exposure to asbestos” is a spatially and temporally distributed event with very fuzzy boundaries, and the onset of cancer a very complicated process involving multiple events or “hits”: initial DNA damage, failure of DNA repair mechanisms and of various tumor suppressor genes, and so on. The complexity of the neurobiological processes underlying having a desire to drink and, one suspects, the absence of any clear answer to the question of exactly which of these are to be “identified” with the desire is, if anything, even more obvious. Similarly for the physical and biological processes underlying the occurrence of inflation.

There are, no doubt, many other possible ways of unpacking what it is for a law (or laws) to underlie a causal claim. Rather than exploring these, let me just note that there is at present no generally accepted account of what the “underlying” relationship involves and, in fact, good reason to suspect that there may be no single (informative and nontrivial) relationship of “underlying” that holds for all causal claims and the laws associated with them. In the absence of such an account, both the underlying and meaning theses are unclear. The reader who thinks that clarification would be easy to provide is invited, as an exercise, to provide it for claims (4.4.1)–(4.4.3).

There is yet another respect in which the meaning thesis is problematic. The thesis claims that all causal claims “entail” in virtue of their meaning, the holding of some associated underlying law. The holding of the law is thus said to be “necessary” for the causal claim to be true or (within a framework in which meaning is linked closely to truth conditions) part of the “truth conditions” of or a “truth maker” for the claim. However, despite the apparent naturalness of the quoted phrases, it is hard to find a way of interpreting them that makes them plausible. Consider again the causal claim (4.2.4) (“The impact of the knee caused the tipping over of the inkwell”). As noted above, the obvious candidates for the underlying laws will include Newton’s laws of motion and various specialized force laws. In what sense are these laws necessary for (4.2.4) to be true? After all, there is a range of possible worlds in which the underlying laws deviate from Newton’s in various small ways and in which (4.2.4) would still be true. It is true that if the underlying laws were *sufficiently* different from Newton’s laws, (4.2.4) would have been false, but of course this does not mean that every world in which Newton’s laws are false is a world in which (4.2.4) is false. In view of this, it is not obvious in what sense (4.2.4) entails the existence of any specific underlying set of laws, Newton’s or any other. If anything of this sort is entailed by (4.2.4), it looks like it is simply that the underlying laws linking the events in (4.2.4), whatever they may be, must fall within the range of possibilities that are consistent with (4.2.4)—a rather deflationary conclusion. This sort of claim is an odd candidate for a “truth maker” for (4.2.4).

I suspect that part of the problem here is that two quite different ideas have been confused: the idea that  $p$  figures in an explanation of  $q$  and the idea that the truth of  $p$  is necessary for  $q$  to be true. The most natural and plausible way of understanding the claim that the laws of, for example, Newtonian mechanics “underlie” (4.2.4) is simply that these are laws to which we might appeal to explain why (4.2.4) holds: the relationship of “underlying” between (4.2.4) and those laws is simply the relationship between explanans and explanandum. However, there is nothing in the idea that Newton’s laws figure in an explanation of (4.2.4) that licenses us to take the holding of those laws as necessary for (4.2.4) to be true or as implied by (4.2.4).

There are also problems with another tempting line of thought that seems to influence Hempel and that might seem to motivate some version of the meaning thesis. This is the thought that the occurrence of  $c$  and  $e$  and the holding of some law appropriately linking them (or  $c$ ’s having property  $P$  and

*e*'s having property *Q* and the holding of some law linking these properties) is at least *sufficient* for "*c* caused *e*" to be true, and arguably the only candidate for such a sufficient condition that is clear and noncircular. The counterexamples (4.2.1)–(4.2.2) to the sufficiency of the *DN* account also show that, at least on any straightforward understanding of "instantiation," the instantiation of a law linking *c* to *e* is *not* sufficient for *c* to cause *e*: the length of the shadow and the height of the flagpole "instantiate" a law linking the two, but the former doesn't cause the latter.<sup>12</sup> Contrary to what is commonly supposed (see, e.g., Kim 1999, p. 17), one cannot argue that (4.2.4) has an implicit *DN* structure on the grounds that the instantiation of such a structure provides a guarantee that (4.2.4) is true that would otherwise be lacking. Nor does the existence of such a law by itself ensure that the knee movement is causally or explanatorily relevant to the tipping of the inkwell. Again, this is not to deny that there are facts about laws and initial conditions that are sufficient for (in the sense of explaining why) claims like (4.2.4) to be (are) true, but whatever these facts are, their relationship to (4.2.4) will be far more complex and indirect than the "instantiation of a *DN* structure" idea suggests.

Skepticism about the meaning thesis is reinforced by another set of considerations. As already observed, the notion of a law of nature, at least in the restricted sense in which it is understood in contemporary science and philosophy, is a relatively recent historical development, the invention and understanding of which requires a considerable degree of intellectual sophistication. As Scriven observes, people made singular-causal claims and offered them as explanations long before anyone thought of the idea of a law of nature. Indeed, it is plausible that many users of causal explanations (singular and otherwise) today do not possess these ideas. The idea that all causal claims must be "backed" by laws is, if anything, even less familiar to users of those claims. In light of these facts, the contention that a claim about the existence of some law of nature is somehow built into or guaranteed by the meaning of singular-causal explanations like (4.2.4) or type claims like "Impacts of knees cause spills" seems dubious. Do users of causal claims who are unfamiliar with the notion of a law of nature (or who live in cultures that are unfamiliar with this notion) operate with a different sense of "cause" or causal explanation than those who are familiar with the notion of a law of nature? Do causal locutions somehow imply facts about the existence of underlying laws even though not only individual users of such locutions but the entire linguistic communities of which they are a part are unaware that this is the case? It is true that there are semantic theories according to which it might be argued that it is possible for the meaning of causal claims to outrun what users of those claims know or understand in something like this fashion.<sup>13</sup> Perhaps, if we had convincing positive reasons for accepting the meaning thesis (and a clear understanding of what "backed by a law" involves), we might appeal to these semantic theories to reconcile the thesis with the fact that the notion of cause antedates the notion of law by thousands of years. But in the absence of such positive reasons, the considerations rehearsed in this section seem to further undermine the thesis.

## 4.5 The Manipulationist Alternative

These doubts are further deepened when we confront the meaning thesis with the manipulationist account of causal claims developed in chapters 2 and 3. As explained there, the references in the formulation of a manipulationist theory **M** and **TC** to what would happen to *Y* under manipulations of *X* should be interpreted as generalizations, the truth of which requires some degree of reproducibility in the response of *Y* to manipulation of *X*. In this sense, any type-level causal claim implies that certain associated generalizations are true. Thus, for example, on a manipulationist theory, the claim that smoking causes lung cancer does imply a generalization to the effect that there is some range of interventions and background circumstances such that when such interventions change whether subjects smoke in these background circumstances, this is reproducibly associated with changes in (or changes in the probability of) whether they develop lung cancer. However, as also explained in chapter 3, the generalizations that follow from the manipulationist conception may (and typically will) lack many of the features standardly assigned to laws. They may be exception-ridden, vague, limited to various special circumstances that are not characterized precisely, and so on. This is reflected in what we think people who are competent users of a generalization like “Smoking causes lung cancer” believe or understand. It is plausible that someone who understands this generalization and asserts that it is true must also have the idea that changing whether people smoke is a way of changing whether they will develop (or the chances of their developing) lung cancer, but not plausible that he must also have the idea that there is a law of nature (in the strict sense) linking smoking to lung cancer.

I take these observations to undermine the meaning thesis but not the underlying thesis. That is, although the manipulationist account casts doubt on the idea that causal claims entail, in virtue of their meaning, the existence of underlying laws, it does not rule out the possibility that, as a matter of empirical fact, such laws exist for most or all causal claims. Although there is nothing in a manipulationist account of causation that implies that a law must underlie every true causal claim, there is also nothing in that account that implies that this claim about nomological backing must be false. However, on a manipulationist account of causation, if there are laws of nature underlying every true causal claim, this will be a further (empirical) fact, and not something that follows just from the meaning of causal claims.

In my view, it is a virtue of a manipulationist account of causation that it does not (or at least need not be understood as) build(ing) strong and demanding claims about the existence of laws into the content or meaning of causal claims. The manipulationist account does not hold the meaning of causal claims hostage to poorly understood ideas about laws “backing” or “making true” causal claims. The manipulationist account also provides a natural explanation of the anthropological fact, noted above, that virtually all human beings make and understand causal claims but few have possessed anything like the notion of a law of nature or the idea that the latter is required to “back”

the former. On the manipulationist account, the content of causal claims is rooted in what we know about changing and manipulating nature, knowledge that can be grasped independently of the notion of a law of nature. My suggestion, in other words, is that preliterate or scientifically unsophisticated people grasp such manipulationist ideas (e.g., that bumping a table hard is a way of spilling a container of liquid resting on it), and this is enough for them to understand the content of causal claims (e.g., that such bumps cause spillings) even if they are entirely innocent of the notion of a law of nature. The fact that the content of causal claims can be independently understood in terms of manipulationist ideas explains why the thesis that causal claims entail, in virtue of their meaning, facts about the existence of laws, is false. It also undercuts Hempel's claim (section 4.1) that if the meaning of claims that contain words like "cause" is to be made clear, such claims must be elucidated in noncausal terms, and that the only plausible way of doing this will take them as having to do with the holding of regularities.

#### 4.6 Laws as Underlying Causal Claims

I said above that although the meaning thesis is highly dubious, it might nonetheless be true that, under some suitable interpretation, the underlying thesis is correct. In my view, the most compelling reason to adopt the underlying thesis does not have to do with metaphysical considerations about the need for nomological "truth makers" for causal claims. Rather, as intimated above, it is simply this: it has become part of our general view of the world that we expect there to be deeper scientific explanations, which appeal explicitly to laws, for why many garden-variety causal claims (such as (4.2.3)–(4.2.6)) are true. The existence of such deeper explanations is not entailed by the meaning of causal claims, as the meaning thesis contends, but it is nonetheless true that they have often been found to exist. Moreover (what is a different matter), there are no cases of true causal claims for which such an underlying law is known not to exist.

In this view of the matter, the "backing" relationship between causal claims and underlying laws is simply the relationship of explanation, and thus no different in principle from the relationship between the claims of some "upper-level" theory (e.g., phenomenological thermodynamics) and another theory (e.g., statistical mechanics) that explains these. When it is understood in this way, we need not endow the backing relationship with other, more metaphysically ambitious (and dubious) characteristics. For example, we need not claim that the holding of underlying laws is "necessary" for causal claims to be true or that the laws are what "make" the causal claims true—ideas that led us into difficulties in section 4.4. We may further agree that causal claims that are straightforwardly inconsistent with known physical laws or scientific fact (e.g., causal claims about telepathy or about the effects of faster-than-light travel) must be false.

Finally, this understanding of what is plausible in the underlying thesis suggests something about what its limitations may be. As remarked above, it

seems unlikely that there are cases in which a causal claim is true and yet we will be able to establish the nonexistence of any law-based explanation for why it is true. In this sense, it seems unlikely that we will ever find cases in which the underlying thesis is clearly false. What seems much more likely is that there are many causal claims for which it is unclear what the backing laws are or even what the backing relation amounts to. These will be cases—(4.41)–(4.43) are perhaps examples—in which the relationship between the causal claim and plausible candidates for underlying laws (and initial conditions) is highly complex and indirect, in which the categories and ontology deployed by the causal claim seem vague and unclear from the perspective of the underlying theory and fail to mesh with it, and in which these problems as well as computational intractabilities and other sorts of epistemological difficulties make the actual exhibition of even the broad outlines of any explanatory relationship impossible. In such cases, it may be harmless to say that an explanation in terms of underlying laws must be “there” even if we do not know what it is, but we should also realize that we do not have much purchase on what “underlying” means. In such cases, the underlying thesis may express little more than a commitment to physicalism and to the idea that physical phenomena are law-governed.

#### 4.7 The Epistemological Thesis

What about the epistemological thesis? Here, I can be briefer. The epistemological thesis is clearly false: there are a variety of procedures, evidence, and patterns of argument (including experimentation and the causal modeling techniques described in chapter 7) that can be used to reliably establish that various sorts of causal claims are true without “relying on” or making explicit use of laws. Scriven is quite right in claiming, against Hempel, that one may establish that (4.2.4) (“The impact of the knee caused the tipping of the inkwell”) is true without making use of information about physical laws. Indeed, if one agrees that the laws associated with many causal claims will be indirect generalizations, the falsity of the epistemological thesis seems to follow almost inevitably, because it is clear that many such laws will be unknown to all or most users of such claims. If this is so, those users cannot be relying on their knowledge of such laws to establish the causal claims in question.

It is also clear that the epistemological thesis does not entail the explanation thesis: even if it were true that, to establish a causal claim, one always had to be able to produce some backing law as evidence, it would not follow that this law was part of the explanation provided by the causal claim or essential to its explanatory import. We do not in general suppose that all of the evidence that supports an explanation is itself part of the explanation. For example, the General Theory of Relativity (GTR) can be used to explain the advance of the perihelion of Mercury and the bending of the starlight by the Sun. The photographic plates taken by the Eddington eclipse expedition of 1919 are part of the evidence for GTR, but the plates are not part of the

explanation of Mercury's behavior. Thus, even if the epistemological thesis were true, some independent argument would be required to show that laws are part of every explanation.<sup>14</sup>

#### 4.8 The Explanation Thesis

I suggested above that if we understand the underlying thesis as the claim that associated with every causal claim is an explanation of that claim that appeals to a law, then although it may be unclear what the underlying thesis requires in connection with some causal claims, it is nonetheless plausible in connection with many others. Suppose, for the sake of argument, that the underlying thesis is true for all or at least many causal claims. Could we then appeal to it to argue that the laws underlying a causal claim are part of (or in some way "involved in" or "invoked by") the explanation provided by the claim, and hence (on the assumption that explanations must provide information about causes) that the explanation thesis is true after all?

Let me begin by noting that even those who accept some version of the explanation thesis recognize that there must be *some* distinction between the laws or argument structures underlying a causal claim and what belongs to the explanation provided by the causal claim. No one supposes that *all* of the explanatory information conveyed by the former is automatically conveyed by the original causal claim itself. Even if we hold that underlying (4.2.4) ("The impact of the knee caused the tipping of the inkwell") is some explanation that appeals to the laws of Newtonian mechanics, it is obvious that (4.2.4) does not convey all of the relevant details of the Newtonian explanation. In this sense, any remotely plausible account of explanation must recognize some distinction between the explanation thesis and the underlying thesis.

I take writers like Hempel and Railton to agree. As I interpret them, their position is not that *all* of the information present in the DN structure or ideal explanatory text that (they suppose) underlies causal claims like (4.2.3)–(4.2.6) is conveyed by (4.2.3)–(4.2.6) themselves. Rather, they defend the idea that claims like (4.2.3)–(4.2.6) convey only *some* of the information in this underlying structure, in a partial or imperfect way, and that the explanatory import of claims like (4.2.4)–(4.2.6) somehow derives from this fact. The crucial issue thus becomes whether this idea of (4.2.3)–(4.2.6) explaining by "conveying (partial) information" about an underlying and partly epistemically hidden structure can be developed in a clear and convincing way.

I think there are at least three reasons why this idea should be rejected. The first is methodological: the appeal to hidden structure makes it too easy to protect one's favored theory of explanation from genuine counterexamples. If we find apparent examples of explanations that do not satisfy the requirements of our favorite theory, we have only to claim that underlying those explanations are other, ideal explanations that do satisfy those requirements, that the original examples explain in virtue of conveying information about the latter, and hence that our favorite theory is vindicated after all. This strategy will



always be available as long as there are reasons to think that there are underlying explanations satisfying the requirements of our favorite theory, even if these requirements are not remotely plausible candidates for necessary conditions on explanation. Suppose that your theory claims that all successful explanations must appeal to second-order differential equations or must be Lorentz-invariant or must trace continuous processes (in some appropriate sense of “continuous”) or must be framed within the current language of superstring theory, and that, as it happens, there is a fundamental physical theory of everything (in the sense that it explains all behavior of all physical objects) that does satisfy these requirements. Given apparent examples of successful explanations from chemistry, biology, or economics that do not satisfy these requirements, you can always say that underlying them are explanations that do satisfy the requirements, that the upper-level explanations work by conveying information about these underlying explanations, and hence that, despite all appearances to the contrary, the requirements are correct after all. If we do not find this sort of argument convincing when the requirement in question is Lorentz invariance, it is unclear why we should find it convincing when it is claimed that apparent explanations that do not cite laws or unify or trace continuous causal processes nonetheless explain in virtue of conveying information about underlying explanations that do these things. At the very least, we need constraints on when a partial explanation may be said to “convey information about” an underlying explanation that rule out arguments of the sort just described.

However, it is far from obvious how to do this. Peter Railton, who has provided the most detailed and careful version of the ideal text account in the recent literature on explanation, forthrightly acknowledges his inability to make the notion of “provides information about” precise, but he suggests that we must make do with the notion anyway (1981, p. 244). He contends that we should understand this phrase to mean something like “reduces uncertainty about.” Very roughly, his idea is that an explanatory claim provides information about an underlying ideal explanatory text (or an underlying nomothetic structure) if the claim reduces uncertainty about some properties of this underlying structure—if the claim provides information that rules in or rules out various possibilities concerning the characteristics of the structure.

As Railton recognizes, this proposal has many counterintuitive consequences. Suppose it is true that (*N*) the laws underlying (4.2.4) (“The impact of the knee . . .”) are those of Newtonian mechanics. This observation certainly provides information in the “reduces uncertainty” sense about the *DN* explanation (or ideal explanatory text) underlying (4.2.4). But if one wants to know why the inkwell tipped over, the observation (*N*) by itself would be regarded by most people as no explanation at all.<sup>15</sup> Similarly, the observation (*D*) that the laws underlying (4.2.4) are deterministic (rather than probabilistic) certainly provides information about the ideal text explanation underlying (4.2.4). Nonetheless, most of us again would judge that (*D*) by itself is no explanation at all of why the inkwell tipped over. Next, consider the claim that (4.8.1) the price of tea in China in 1835 is not part of (plays no role in) the *DN* structure

of or any other ideal explanatory text that underlies (4.2.4). (4.8.1) is presumably true and, in the relevant sense, conveys information about the ideal text underlying (4.2.4). Again (4.8.1) does not explain (is not even part of an explanation of) why the inkwell tipped over. Indeed, paradoxically, on Railton's account, even the claim (assuming that it is true) that there are no causes for an outcome, or no factors on which it depends, furnishes information about the ideal explanatory text for the outcome and hence is explanatory with respect to the outcome. Moreover, as Railton himself notes, the following counts as an "explanation" for an episode of radioactive decay: "The relevant ideal text contains more than  $10^2$  words in English" (1981, p. 246).

In connection with this last case, Railton suggests that we need to distinguish between a poor explanation and no explanation at all. The information about number of words is explanatory but only minimally so; it is at "the very low end of the continuum of explanatoriness" (1981, p. 246). Presumably, Railton would respond similarly to the other examples described above. This response seems unsatisfactory for several reasons. First, anyone who is not already in the grip of a theory like Railton's would judge claims like (N), (D), and (4.8.1) to be no explanations at all, rather than poor explanations. Second, and more important, our judgments about when explanations are more or less good don't seem to line up with the amount of information they provide about ideal texts in the way that Railton's account requires. The information that (N) the laws underlying (4.2.4) are those of Newtonian mechanics is, according both to the *DN* model and Railton's nomothetic model, of central importance to successful explanation. Yet we simply don't think of it as answering the question "Why did the inkwell tip over?" By contrast, the information that I hit the desk with my knee does seem to explain the tipping over, even though it is difficult to see in what sense it furnishes more information about the ideal text associated with (4.2.4) than (N) does. Similarly, in a theory like Railton's, it appears that the information that there are no causes for an outcome is not merely explanatory, but maximally so—because this conveys all of the information in the relevant explanatory text. Needless to say, this does not coincide with the judgment of most people who are not already committed to Railton's theory. Again, the goodness of this explanation does not appear to be just a function of how much it reduces uncertainty about what is in the associated ideal explanatory text.

Many of the cases just mentioned are uninteresting from the point of view of practical methodology, in the sense that everyone agrees that they are poor or no explanations, even though it may be unclear how to get the ideal text account to yield this conclusion. However, there are other cases in which the ideal text approach has methodological implications that are both nontrivial and misguided. Consider what Railton calls "explanation by correlation." He writes:

To this species [of statistical explanation] belong the many efforts to use the direct results of multivariate analysis and other statistical techniques as explanations of behavior of (often complex) causal systems. It is

unnecessary in this connection to delve into the details of explanations of this kind, for the point to be made is highly general: the nomothetic account enables us to see how statements of statistical correlations may function not only as evidence for causal connections, but also as sources of information about aspects of ideal causal explanatory texts. Thus a statement that there is a “statistically significant” correlation between exposure to high levels of cotton fiber in the air and incidence of brown lung disease may be put forward to explain why textile workers in cotton mills experience abnormally high rates of brown lung. Here the explanation may involve an irreducibly probabilistic process, but it need not. Let us suppose for the sake of illustration that it does not; still, the statement of correlation may convey information about the relevant causal ideal text for explaining the frequency of brown lung among cotton mill workers, since it points to a substance, cotton fiber, the presence of which in the air breathed is an important element of this text. In this light, claims about the “strength” of a correlation, or about the amount of variation in one variable that can be “statistically explained” in terms of the variation of another, may be seen as ways of providing information about the extent and independence of the roles of various factors in relevant ideal explanatory texts. (1981, pp. 252–53)

The issue of when a regression equation and other results of multivariate analysis are explanatory is a live and important methodological problem, which is discussed in detail in chapter 7. The results of that discussion were anticipated in chapter 2: correlations, in themselves, are never explanatory. What is required for a variable  $X$  to be explanatory with respect to some other variable  $Y$  is that there be interventions on  $X$  that will change the value (or probability distribution) of  $Y$  or, equivalently, that the relationship between  $X$  and  $Y$ , as specified, say, in a regression equation linking  $X$  to  $Y$ , be invariant under some range of interventions on  $X$  that will change  $Y$ . Thus, for example, if  $X$  and  $Y$  are correlated but  $X$  does not cause  $Y$ , the correlation instead being due to  $Y$ 's causing  $X$  or to the presence of a common cause  $C$  for both  $X$  and  $Y$ , then  $X$  will not figure in an explanation of  $Y$ .

This conclusion is generally accepted by methodologically sophisticated social scientists but often not respected in practice. Social science (as well as psychology and epidemiology) journals are full of studies in which the presence of “statistically significant” correlations between  $X$  and  $Y$  (perhaps when certain other variables are controlled for) is taken to establish that  $X$ s cause or explain  $Y$ s, without any independent argument that changing  $X$  will change  $Y$  or any argument that rules out the possibility that the correlation between  $X$  and  $Y$  is due to some other source, such as the presence of a common cause. Railton's remarks about explanation by correlation appear to endorse this misguided practice. Moreover, this is not an isolated slip. The conclusion that citing a correlation is in itself explanatory follows directly from the ideal text account of explanation. Citing the existence of a correlation between  $X$  and  $Y$  does indeed convey information (in the diminishment of uncertainty sense) about the contents of the ideal explanatory text describing whatever underlies the correlation. If there is a correlation between exposure to cotton fiber and

brown lung, then we have (at least) good reason to think that either the former causes the latter or vice versa or both have a common cause or causes. All alternative accounts according to which neither of these three possibilities obtain are made much less plausible.<sup>16</sup> However, if the arguments of chapters 2 and 7 are correct, the fact that the ideal text account implies that citing correlations is explanatory is a defect rather than a virtue. It provides yet another illustration of the point that the ideal text account is too permissive: the citing of a correlation can convey information about an underlying correct explanation without itself being explanatory.

Railton's discussion and, in particular, his remark that "it is unnecessary . . . to delve into the details of explanation of this kind" also illustrate another unfortunate consequence of ideal text accounts. By viewing various nonideal forms of explanation (such as that provided by regression under certain conditions) as of interest only insofar as they serve as (highly imperfect) vehicles for conveying information about underlying explanations of a quite different structure, the ideal text approach encourages us to overlook the need for a detailed analysis of how such nonideal explanations actually work and of the features that distinguish those attempts at nonideal explanation that are genuinely explanatory from those that are not. We stop far too soon in our analysis of such explanations if we say only that they are explanatory in virtue of conveying information about an underlying ideal text.

#### 4.9 Explanation and Epistemic Accessibility

There is yet another reason for rejecting the hidden structure strategy. This derives from a general point about the epistemology of explanation and the connection between explanation and understanding, a point that has application well beyond the details of disputes about the proper analysis of claims like (4.2.3)–(4.2.6). It is a plausible constraint on what an explanation is that it must be something that provides understanding. To say that certain information is "part" of an explanation or contributes to its explanatory import is to say that this information contributes to the understanding provided by the explanation. This in turn imposes an epistemic constraint on what information can be part of an explanation and can contribute to its explanatory import: such information must be epistemically accessible to those who use the explanation. Put slightly differently, the idea is that the features of the explanation that endow it with explanatory import—that make it an explanation—must be features that can be known or grasped or recognized by those who use the explanation; if not, it isn't in virtue of possessing those features that the explanation produces understanding. On this way of looking at matters, there is something deeply puzzling about the suggestion that claims like (4.2.3)–(4.2.6) explain or convey understanding in virtue of providing information about the existence of some underlying epistemically hidden structure, whether a *DN* argument or an ideal explanatory text. The mere obtaining of this structure, independently of anyone's awareness of its existence, cannot be

what accounts for people's judgment that, for example, the impact of the knee on the desk is explanatorily relevant to the tipping over of the inkwell. If this line of thought is correct, it seems to follow that to the extent that information about the laws or structures that underlie singular-causal (or other sorts of causal) claims is epistemically hidden from those who use such explanations, it cannot be that this information contributes to the explanatory import of these explanations.<sup>17</sup>

With this idea about epistemic accessibility in mind, let us return to (4.2.3) ("Jones's latent syphilis caused his paresis"). Contrast two situations. In the first, there exists an additional condition *K*, which, in conjunction with having untreated latent syphilis, is nomologically sufficient for having paresis, but no one knows this or has any information about this condition. In the second situation, development of paresis is an irreducibly indeterministic process. If one has untreated syphilis, one has a probability of 0.25 of contacting paresis, and no further factors are relevant to this probability. Again, however, no one knows that this is the case. My judgment is that (4.2.3) is just as good or as bad an explanation of why Jones developed paresis in both of these situations. My claim about the connection between explanatory information and epistemic accessibility supports this judgment: the information about what paresis depends on that is conveyed by (4.2.3) is the same in both cases, and because the additional facts about whether or not *K* obtains are epistemically inaccessible in both situations, they contribute nothing to the explanatory import of (4.2.3). By contrast, someone who thinks that the explanatory import of (4.2.3) somehow depends on its conveying information about an underlying *DN* or *IS* structure will presumably regard (4.2.3) as explanatory in the first situation but not the second—a judgment that seems to me to be highly counterintuitive.

This is not to say that it is never correct to think of explanations as having an implicit structure or content in addition to what is conveyed in their explicit verbal or mathematical formulation. In fact, I believe that such nonexplicit structure is present in many explanations. Indeed, in the next chapter, I develop an account of a kind of explanation that figures in many areas of science that has just this feature. According to this account, various sorts of background information that may not be explicitly represented in the usual formulation of such explanations make an essential contribution to their explanatory import. However, the epistemic condition introduced above constrains what this background information can be like: it must be information that those who use such explanations know or can reasonably be taken to believe and that they recognize as relevant to the import of the explanation they offer. Similarly for explanations such as (4.2.4)–(4.2.6): there is nothing wrong with the idea that such explanations have implicit features that are relevant to their explanatory import, but again, these must be epistemically accessible features.<sup>18</sup>

The following chapters explore the implications of these points about epistemic accessibility and the inadequacy of hidden structure and ideal text accounts for the treatment of a number of explanations drawn from biology and the social and behavioral sciences that do not appear to explicitly cite

laws (at least, if we have even a modestly demanding conception of what a law is). Many philosophers are attracted to accounts of such examples, according to which, although they do not explicitly cite laws, we should nonetheless think of them as explaining in virtue of “providing information” about some underlying law, even if this law is unknown to (and perhaps for all practical purposes unknowable by) all who use the explanation, and even if the very fact that there is some such (unknown) law is unrecognized or regarded as irrelevant by users of the explanation. In this way, they attempt to reconcile the failure of such explanations to explicitly cite laws with the claim that laws are nonetheless “required for” or “involved in” all successful explanations. The arguments given above are meant to suggest that this general strategy is illegitimate: it is not an adequate account of how such explanations work to simply say that they “provide information about” an underlying explanatory structure, because a claim can provide information about such a structure without itself being explanatory. Moreover, considerations having to do with epistemic accessibility constrain the “implicit structure” we can plausibly attribute to such explanations. More generally, to understand how explanations like (4.2.3)–(4.2.6) work, we need to see them as explanatory structures in their own right and not merely as vehicles for conveying information about underlying explanations of a quite different form. We need to identify structural features they possess that are epistemically accessible and whose presence or absence is relevant to whether they are good or bad explanations. At least in their present forms, ideal text and hidden structure accounts do not help us with these tasks.

#### 4.10 Interim Conclusions

What may we conclude from this long discussion of various counterexamples to the *DN/IS* model? I suggest that several conclusions follow. First, we may conclude from (4.2.1) and (4.2.2) that the *DN/IS* model does not state sufficient conditions for successful explanation. Second, we may conclude from (4.2.3)–(4.2.6) and the failure of various versions of the hidden structure strategy that the *DN/IS* model does not state necessary conditions for successful explanation. Next, recall the distinction between the *DN/IS* idea that explanation is matter of nomic expectability and the broader idea that explanation is “nomothetic” in the sense that it is a matter of nomic subsumption; that is, that it is necessary and/or sufficient for *C* to explain (or figure in an explanation of or be explanatorily relevant to) *E* that *C* be shown to linked by a (deterministic or probabilistic) law to *E*, or that there be a law that “subsumes” *C* and *E* or that *C* and *E* “instantiate.” The nomic subsumption conception of explanation is also misguided. (4.2.1) and (4.2.2) show that *C* may be linked by a law to *E* and yet *C* may be explanatorily irrelevant to *E*. (4.2.3)–(4.2.6) show that *C* may be explanatorily relevant to *E* even though no law linking *C* to *E* is explicitly exhibited and even though it is wrong to think of any such law as being part of the explanation in which *C* figures. As will become apparent in

chapter 5, this is *not* to deny that laws play an important role in (some) explanations. Rather, the point is that both the nomic expectability and, more broadly, the nomic subsumption account of how laws are relevant to successful explanation is mistaken.

#### 4.11 Explanation in the Absence of Clear Criteria for Lawhood

I turn now to another set of difficulties with the *DN* and other nomothetic models, again having to do with the role they assign to laws in explanation. These derive from the fact, noted in section 4.3, that there is no generally accepted account of laws themselves, either within the literature on explanation or, for that matter, elsewhere within philosophy of science. If we don't know what laws are, or what sorts of conditions a generalization must satisfy in order to qualify as a law, how can we get a grip on what laws contribute to (or whether they are required for) successful explanation? In his discussion of this issue, Hempel (1965a, pp. 338ff) considers a number of standard criteria that have been proposed to distinguish between laws and accidental regularities—these largely coincide with those described in section 4.3—and concludes that no existing accounts of lawhood are fully successful. Some of the proposed criteria do not seem to be necessary for lawhood, others are satisfied by both laws and accidental regularities, and still other criteria are unclear or “circular.” Hempel concludes that the problem of finding necessary and sufficient conditions for lawhood has proved “highly recalcitrant” (p. 338).

Writing about this issue nearly thirty-five years later in his well-known survey, Wesley Salmon (1989) reaches a similar conclusion. He remarks that the “problem of distinguishing laws from accidental generalizations” has proved “extraordinarily difficult” and that no one has successfully explicated the distinction (pp. 15ff). Indeed, Salmon lists the solution to this problem as the first item of business in the “Agenda for the Fifth Decade” (of work on explanation), with which he concludes his survey.

Adherents of the *DN* and other nomothetic models are thus in the uncomfortable position of insisting that the distinction between laws and accidental generalizations is crucial to explanation, even though there appears to be no satisfactory account of the distinction. (Similarly, philosophers who hold that causal claims are unclear in meaning unless cashed out in terms of facts about the obtaining of laws are in the odd position of holding that considerations of clarity require an analysis of causation in terms of a notion that no one has been able to make clear.) Moreover, the unclarity of the notion of law is not merely an abstract or philosophical difficulty for nomothetic models, but affects the practical applicability or usefulness of such models. Some writers deny that there are any laws at all (a view I discuss in chapter 6). Putting aside such views for the moment and granting that some laws exist, the clearest examples seem to be found in sciences like physics and chemistry. By contrast, uncontroversial examples of laws are less easy to find in sciences like biology and

geology and harder still to find in the social and behavioral sciences. Many explanatory generalizations in such disciplines appear to lack features commonly assigned to laws; for example, they may hold only in limited domains, have a rather open-ended set of exceptions that defies any simple characterization, and may be imprecise and vague. Both some practitioners and some philosophical commentators do indeed use the honorific “law” to describe certain generalizations in the special sciences (Mendel’s “laws,” the “law” of supply and demand, the “law” of effect, etc.), but many other practitioners and commentators deny the appropriateness of this word. In the special sciences, one does find a contrast between descriptive and explanatory generalizations, but, rather than being described as laws, the latter are often described as *principles* or as generalizations describing the operation of *mechanisms* or causal relationships.

When we attempt to apply the *DN* model (or, for that matter, any nomothetic model of explanation) to areas of science outside of physics and chemistry, we thus encounter a very tangled dialectical situation. One possible view of the matter is that explanation does indeed require the explicit citing of laws (this, after all, is what a straightforward reading of the *DN* and nomothetic models without “hidden structure” or “underlying ideal text” embellishments seems to entail) and that, because the special sciences contain few if any laws, they are largely or entirely unexplanatory. Some writers do embrace this conclusion or something close to it (see, e.g., Rosenberg 1994), but it has seemed arbitrary and high-handed to most commentators. The more common response to the apparent paucity of laws outside of physics and chemistry has been either (a) to opt for a watered-down view of what a law is so that generalizations that are vague, gappy, have exceptions, or hold only in limited domains can nonetheless count as laws (the absence of generally agreed on criteria for lawhood facilitates this strategy), and hence have explanatory import, and/or (b) to appeal to the hidden structure strategy to claim that the generalizations of the special sciences, although perhaps not themselves laws, nonetheless have explanatory import in virtue of “tacitly invoking” or conveying information about an unknown underlying structure that does involve laws. Typically, when adherents of (a) are pressed to explain how gappy generalizations can be used to explain, they advert to (b).

Position (b) suffers both from the deficiencies of the hidden structure strategy and from the lack of clear criteria for what laws are, as well as the absence of a convincing account of the role of laws in explanation. The position I defend in subsequent chapters avoids these difficulties. In my view, generalizations in the special sciences can be used to explain, as long as they are invariant in the right way, whether or not they are regarded as laws. It is not only true that the traditional criteria for lawhood are of limited helpfulness in distinguishing laws from accidental generalizations. It is also true that whether or not a generalization is explanatory (or invariant) is surprisingly independent of whether it satisfies these criteria. Because invariance is the key to explanatory power, we don’t need to decide whether a generalization counts as a law (and hence we don’t need to find a sharp dividing line between laws and nonlaws) to



distinguish between the explanatory and the nonexplanatory. The features that distinguish explanatory from nonexplanatory generalizations are not the same features that distinguish laws from nonlaws. Such a position requires not just a rejection of the *DN* intuition that explanation is a matter of providing nomic grounds for expecting, but also a rejection of the more general idea that explanation is always a matter of subsumption under laws. I attempt to articulate an alternative conception of explanation that avoids these commitments in the following chapters.

## 4.12 What's Right about the *DN* Model

Despite the limitations in the *DN* model described above, it possesses a number of other features that are attractive and that I believe should be retained in any viable theory of explanation.

### 4.12.1 *Objectivity*

The first is that the *DN* model makes whether an explanans explains an explanandum an "objective" or "nonrelative" matter, independent of the idiosyncratic interests or background beliefs of particular explainers or their audiences. In the case of the *DN* model, these aspirations to objectivity are in part cashed out in terms of the idea that a certain logical relation must obtain between the explanans and explanandum. Quite apart from the considerations described above, there are independent reasons, described in chapter 5, why this particular approach will not work: we cannot capture the notion of explanatory relevance in terms of formal logic alone, as the *DN* model aspires to do. However, the broader goal of trying to construct a theory of explanation that focuses on objective, interest-independent relationships connecting explanans and explanandum and of trying to find a common structure to such relationships that holds across explanations drawn from many different areas of science is very much worth aiming at.

### 4.12.2 *Unified Treatment*

The *DN* model also has the great advantage of offering a unified treatment of two apparently different kinds of explanation: singular-causal explanation and so-called scientific or theoretical explanations. Rather than claiming that these are two completely distinct genres of explanation that work in entirely different ways and have nothing in common, the *DN* model purports to describe a common structural feature that both share, a feature that makes it legitimate to regard both as explanations. According to the *DN* model, this common structural feature is that both kinds of explanation work by showing that their explananda are nomically expectable, although in the case of singular-causal explanation, this is only implicit. I have argued that this particular diagnosis of what the common structural feature consists in is implausible. Nonetheless,

the goal of finding some such common structural feature that is shared by many different varieties of explanation is a highly desirable one. The theory I propose below also attempts to do this. I suggest that there is a common structural feature shared by both singular-causal and scientific explanations which entitles us to take both as explanatory, but argue that this is a different structural feature from the feature on which *DN* theorists focus.

#### 4.12.3 *Fit with Science*

As we have seen, adherents have attempted to motivate the *DN* model by appeal to a certain conception of what scientific understanding consists in and by appeal to Humean ideas about the connection between causal claims and regularities. I have argued that both motivations are unconvincing. However, there is another, more general source of appeal of the *DN* model that has received surprisingly little emphasis in the philosophical literature.

It is an undeniable fact that many explanations in both physics and chemistry, and in other sciences like economics and evolutionary biology as well, in which there is extensive reliance on mathematical theory, involve writing down systems of equations and solving them or constructing derivations from them to relevant features of the phenomenon one wants to explain. Moreover, many of these equations are laws or at least fundamental explanatory principles of considerable generality. Thus, explaining various macroscopic electromagnetic phenomena will typically involve writing down and solving one or more of Maxwell's equations; explaining some elementary quantum-mechanical phenomenon will involve modeling some physical system (and choosing a Hamiltonian for it) in such a way that we can apply the Schrödinger equation to it, then actually solving the Schrödinger equation for the system; and explaining the behavior of a firm will involve the use of information about the marginal cost and revenue curves facing the firm and the general explanatory principle that the firm will act so as to maximize profits to construct a derivation of the behavior of the firm. More generally, although it is not true that *all* explanations involve the construction of explicit derivations from assumptions that include laws or general principles, the use of such derivations is a pervasive feature of explanatory practice in many areas of science, and a feature that any adequate theory of explanation must acknowledge.

Much of the apparent plausibility of the *DN* model derives, in my view, from the fact that it seems to codify and to provide a motivation for such practices. However, this general point about the role of derivations from laws in explanation should be detached from the particular analysis of how such derivations work so as to provide understanding that is put forward by *DN* theorists. Where the *DN* model goes wrong, in my view, is not in its contention that derivation from laws plays an important role in many explanations, but in claiming that *all* explanation must have this structure and that explanation is simply a matter of providing nomic grounds for expecting. In chapter 5, I develop a theory that recognizes that derivations from laws play

an important role in many explanations, but that relies on a different (non-*DN*) account of how such derivations work so as to provide understanding. My suggestion, in other words, is that we can decouple the general observation about explanatory practice just described, which I think is correct, from the *DN* theory of understanding, which is mistaken.

## A Counterfactual Theory of Causal Explanation

### 5.1 Examples and a Contrast

I turn now to the task of constructing a positive theory of explanation. I begin by reminding the reader that the theory that I will be developing is restricted to *causal* explanations. To the extent that there are forms of explanation that are noncausal (in the broad sense of “causal” that concerns me), they will be (aside from a brief diversion in 5.9) outside the scope of my discussion. I often register this restriction in what follows by putting the qualifier “causal” in front of “explanation,” but even when this qualification is not present, the reader should think of it as implicit.

Consider the following deductively valid arguments:

(5.1.1) All ravens are black.

*a* is a raven

*a* is black

(5.1.2) All emeralds are green.

*c* is an emerald

*c* is green

Assume, for the sake of argument, that generalizations in (5.1.1) and (5.1.2) qualify as “laws” and that their other premises are true. Both derivations then satisfy the requirements for an acceptable *DN* explanation.

Let us compare these derivations with two other examples, one from a physics textbook and the other from an economics text. that might more naturally be regarded as typical or paradigmatic cases of theoretical explanation in science.

Consider first an explanation (5.1.3), in terms of Coulomb’s law, of why the magnitude of the electric intensity (force per unit charge) at a perpendicular distance *r* from a very long fine wire with a positive charge uniformly distributed along its length is given by the expression

$$E = \frac{1}{2\pi\epsilon_0} \frac{\lambda}{r}$$

where  $\lambda$  is the charge per unit length on the wire and *E* is at right angles to the wire.

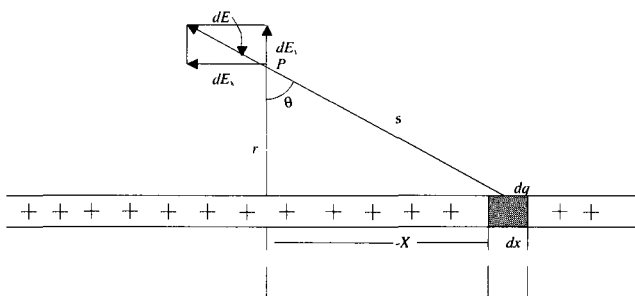


Figure 5.1.1

We can think of the wire as divided into short segments of length  $dx$ , each of which may be treated as a point charge  $dq$  (see figure 5.1.1). The resultant intensity at any point will then be the vector sum of the fields set up by all these point charges. By Coulomb's law, the element  $dq$  will set up a field of magnitude

$$dE = \frac{1}{4\pi\epsilon_0} \frac{dq}{s^2}$$

at a point  $P$  a distance  $s$  from the element. Integrating the  $x$ - and  $y$ -components of  $dE$  separately, we have

$$E_x = \int dE_x = \int dE \sin \theta$$

$$E_y = \int dE_y = \int dE \cos \theta$$

If  $\lambda$  is the charge per unit length along the wire, we have  $dq = \lambda dx$ , and

$$dE = \frac{1}{4\pi\epsilon_0} \frac{\lambda dx}{s^2}$$

The integration will be simplified if we integrate with respect to  $d\theta$  rather than  $dx$ . From figure 5.1.1

$$x = r \tan \theta \text{ and } s = r \sec \theta$$

and thus,

$$dx = r \sec^2 \theta d\theta$$

Making these substitutions, we obtain:

$$E_x = \frac{1}{4\pi\epsilon_0} \frac{\lambda}{r} \int \sin \theta d\theta$$

$$E_y = \frac{1}{4\pi\epsilon_0} \frac{\lambda}{r} \int \cos \theta d\theta$$

If we assume that the wire is infinitely long, the limits of integration will be from  $\theta = \pi/2$  to  $\theta = \pi/2$ . Integrating, we obtain

$$E_x = 0$$

$$E_y = \frac{1}{2\pi\epsilon_0} \frac{\lambda}{r}$$

This shows that the resultant field will be at right angles to the wire and that its intensity is given by

$$E = \frac{1}{2\pi\epsilon_0} \frac{\lambda}{r}$$

As a second example (5.1.4), consider the standard explanation given in microeconomic theory for why a monopoly that takes over a formerly competitive industry will raise prices and restrict output (figure 5.1.2).

When a monopoly takes over a formerly competitive industry, the demand curve of that industry becomes the demand curve, or the average revenue curve (i.e., the curve that gives price per unit at each level of output) for the monopolistic firm. This curve, which is labeled *AR* in figure 5.1.2, will be downward-sloping: price will be inversely related to quantity of product sold. The curve labeled *MR* is the marginal revenue curve (i.e., the curve that gives the change in total revenue occurring with each change in output) for the monopoly. Because the average revenue curve is downward-sloping, this curve will be more sharply downward-sloping. The curve labeled *SMC* is the short-run marginal cost curve for the firm (i.e., the curve that indicates total change in cost for the monopoly per change in output).

Now it is easy to see that if the monopoly maximizes profits, it will select that price  $\bar{P}$  at which marginal revenue is equal to marginal cost. Suppose that the firm had selected price  $P_1$ , at which marginal revenue exceeds marginal cost. In that case, if the firm were to lower its price so that the quantity of goods demanded would rise, revenue would increase at a faster rate than costs, so that profits would rise. Thus, the profit-maximizing firm will lower its price toward  $\bar{P}$ . Suppose, on the other hand, that the firm had selected price  $P_2$ . At

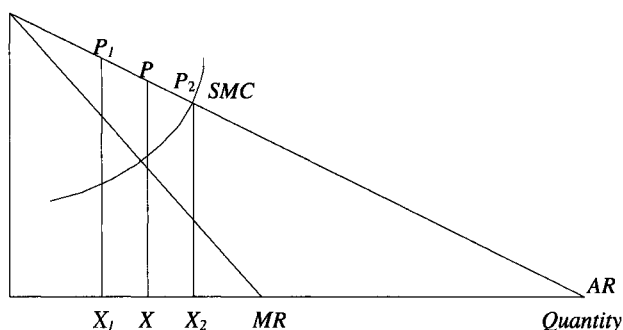


Figure 5.1.2

this price, costs are increasing more quickly than revenue, and profits may be increased by increasing price until  $\bar{P}$  is reached.

Now contrast the behavior of the monopoly with the behavior of a price-taking firm in a competitive industry, before it is taken over by the monopoly. Any such firm will sell goods at  $P_2$ , at which price is equal to marginal cost. Such a firm by definition can sell any amount of its output at the going market price. Its average and marginal revenue curves are identical, and it cannot increase profits by restricting the quantity of goods it sells.

Because marginal revenue is always less than average revenue for the monopolistic firm, marginal cost will always exceed marginal revenue at the price  $P_2$  adopted by the competitive firm. Thus, the monopoly will always be, in comparison with the price-taking firm, a price raiser and output restrictor: it will raise prices from  $P_2$  to  $\bar{P}$  and restrict quantity of goods sold from  $X_2$  to  $\bar{X}$ .

(5.1.3) and (5.1.4) will strike most readers as deeper or more satisfying explanations than (5.1.1) and (5.1.2). What accounts for the differences in explanatory import between these two sets of explanations?

Let us begin by noting that (5.1.3)–(5.1.4), like (5.1.1)–(5.1.2), do indeed exhibit something like the features emphasized by *DN* theorists. In each case, we have a derivation or deductively valid argument, and the conclusion of this argument describes what is being explained. In each case, a generalization of considerable scope and breadth figures essentially in this derivation. In the case of (5.1.3), the relevant generalization (Coulomb's law) is described both by the authors of the text from which the example is taken and by most practicing physicists as a "law." However, we should note for future reference that even in this case, this generalization does not have all of the features commonly ascribed to laws by philosophers. For one thing, there is an obvious sense in which Coulomb's law fails to hold universally. It holds only in a classical regime, in which quantum-mechanical effects are unimportant, and breaks down outside this regime, for example, when the distances between charged particles are sufficiently small. Moreover, the conditions under which Coulomb's law holds are not, at least explicitly, incorporated into formulation of the law on which (5.1.3) relies: the law is stated in an unqualified form that says nothing about the restricted conditions under which it is true. In the case of (5.1.4), we again find essential reliance on a generalization thought to hold in a wide variety of different cases, in this case, the generalization that firms will act so as to maximize profits. However, here too, it is doubtful this generalization has all the features ascribed to laws of nature by philosophers; for example, it seems implausible that it holds universally or without exception.

Despite the fact that we find something not too unlike a *DN* structure in (5.1.3)–(5.1.4) as well as in (5.1.1)–(5.1.2), we think that there are important differences between these two groups of explanations. Outside of philosophical discussions, (5.1.1)–(5.1.2) are not regarded by anyone as serious explanations, or at least they are not advanced as such in scientific textbooks and monographs. (5.1.3)–(5.1.4) seem to possess some feature that (5.1.1)–(5.1.2) lack. What is this feature, and why is it important?

Here is my proposal: explanation is a matter of exhibiting systematic patterns of counterfactual dependence. Not only can the generalizations cited in (5.1.3)–(5.1.4) be used to show that the explananda of (5.1.3)–(5.1.4) were to be expected, given the initial and boundary conditions that actually obtained, but they also can be used to show how these explananda would *change* if these initial and boundary conditions had changed in various ways. (5.1.3)–(5.1.4) locate their explananda within a space of alternative possibilities and show us how which of these alternatives is realized systematically depends on the conditions cited in their explanans. They do this by enabling us to see how, if these initial conditions had been different or had changed in various ways, various of these alternative possibilities would have been realized instead. Put slightly differently, the generalizations cited in (5.1.3)–(5.1.4) are such that they can be used to answer a range of counterfactual questions about the conditions under which their explananda would have been different (*what-if-things-had-been-different* or *w-questions*, for short). In this way, (5.1.3)–(5.1.4) give us a sense of the range of conditions under which their explananda hold and of how, if at all, those explananda would have been different if the conditions in (5.1.3)–(5.1.4) had been different. It is this sort of information that enables us to see that (and how) the conditions cited in the explanans of (5.1.3)–(5.1.4) are explanatorily relevant to these explananda. It is also information that is relevant to the manipulation and control of the phenomena described by these explananda. By contrast, this feature is absent from (5.1.1)–(5.1.2), or at least present only in a greatly attenuated form. As we shall see, the absence of this feature is the common thread running through the counterexamples to the *DN* model described in chapter 4. It is the presence of this feature in (5.1.3)–(5.1.4) and its absence from (5.1.1)–(5.1.2) that accounts for the very different explanatory credentials of these two sets of explanations.

To illustrate in more detail what this feature involves, consider (5.1.3). The generalization—Coulomb's law—employed in (5.1.3) and the overall strategy of integrating over the contributions made by small current elements to get the total field are such that they can be combined with a wide range of different assumptions about initial and boundary conditions to tell us how the explanandum (5.1.3) would have been different had initial and boundary conditions been different in various ways. We can think of this *what-would-happen-if* information as taking several different forms. Some of it is explicit in the expression for the field intensity  $E$ . This expression makes it clear how the field would change if the charge along the wire were increased or decreased or if we were to measure the field at a different distance from the wire. In this way, (5.1.3) exhibits how the field depends on these factors. We learn from (5.1.3), for example, that despite the fact that Coulomb's law is an inverse square law, the field produced by a long straight wire falls off as the reciprocal rather than the squared reciprocal of the distance from the wire.

In addition to this, we can use Coulomb's law and a similar sort of strategy of integrating over the contributions made by small current elements to the total field to show how the expression for the electric field would vary—how



the field would have been different—if the long straight wire in (5.1.3) were instead twisted into a circle of finite diameter or coiled up into a solenoid or somehow deformed into or replaced by a sphere or by two uniform charged plates. For example, in the case of a uniformly charged sphere, we can deduce, again using Coulomb's law, that the field outside the sphere is given by

$$E = \frac{1}{4\pi\epsilon_0} \frac{Q}{r^2}$$

where  $Q$  is the total charge of the sphere (i.e., the field is the same as would be produced by a point charge  $Q$  concentrated at the center of the sphere). Similarly, we can deduce that the field between two charged sheets that are infinite in extent is

$$E = Q/\epsilon_0$$

between the sheets and 0 outside.

More generally, we can see from the above derivations that certain factors (e.g., the geometry of the conductor, the distribution of charge along it, in some cases the distance from the conductor) all make a systematic difference to the intensity and direction of the field. We also can see that various other factors and conditions are irrelevant to the field intensity. For example, we can see from (5.1.3) that the specific material out of which the conductor is made, or its mass, or whether it is constructed by experimenter  $A$  or experimenter  $B$  or set up at spatial location  $X$  or spatial location  $Y$  makes no difference to the direction and intensity of the resulting field. By not mentioning these factors, (5.1.3) tells us that variations in them are irrelevant to its explanandum.

To be sure, the derivation exhibited in (5.1.3) does not by itself tell us what the field produced by a circular wire and so on would be. However, as anyone who has taken a course in classical electromagnetism will attest, one does not learn how to construct a derivation like (5.1.3) or how to solve the problem of the field produced by the wire in (5.1.3) in isolation. Rather, learning classical electromagnetism and coming to appreciate its explanatory resources involves learning how to solve a whole range of problems of the sort described above—that is, learning how to derive an expression for the resulting field for a range of differently shaped and charged conductors in just the fashion I have described. Those who use (5.1.3) typically will be fully aware of these alternative possible derivations; this will not be information that is “epistemically hidden” (in the sense that concerned us in chapter 4) from typical users of (5.1.3). This sort of information about what would happen to the field under a range of different initial conditions is essential to the explanatory import of (5.1.3).

A similar pattern is present in (5.1.4). Whether or not we think of the generalization about profit maximization that figures in (5.1.4) as a law, this generalization can be used to show how changes in the initial conditions cited in (5.1.4) (e.g., changes in the marginal revenue curve or short-run marginal cost curve faced by a firm or more generally in the structure of the market in

which the firm operates) will produce corresponding changes in the behavior of the firm. Thus, we can see from (5.1.4) that if the slope of the firm's average revenue curve were to become less steep, it would lower its prices. Similarly, we can also see from (5.1.4), again by employing the same generalization to the effect that the firm will maximize profits and combining this with information about the average and marginal revenue curves faced by the firm, how the firm would behave differently if it were in a competitive market rather than being a monopoly. We thus see what difference the fact that the firm is monopolistic makes—that this results in an increase in prices and a decrease in output in comparison with an otherwise similar firm in a competitive market.

By contrast, although (5.1.1)–(5.1.2) arguably provide nomic grounds for expecting their explananda, they fail to possess the features described above: they don't answer an appropriate range of what-if-things-had-been-different questions about their explananda and don't locate their explananda within a range of alternative possibilities.

I will focus on (5.1.1) because parallel remarks will also apply to (5.1.2). Unlike (5.1.3)–(5.1.4), (5.1.1) does not identify a condition such that changes in that condition would lead to some alternative to its explanandum, nor does it cite a generalization that can be combined with different initial conditions to yield such alternatives. Put more simply, (5.1.1) doesn't tell us about the conditions under which raven *a* would be some color other than black. (5.1.1) does not lead us to see the actual explanandum (that *a* is black) as one of a range of possible different alternatives, any one of which might have ensued, had initial conditions been different in various ways. Relatedly, (5.1.1) doesn't tell us anything relevant to the manipulation and control of the outcome supposedly being explained; it doesn't tell us anything about how to *change* the color of raven *a* or any other raven or bird. Indeed, in contrast to (5.1.3), where there is a well-defined physical operation of changing the shape of the wire or its charge density (and, in principle, a determinate answer to the question of what would happen to the resulting electromagnetic field if we were to do this), it is unclear, to say the least, what would constitute changing or manipulating the property of being a raven. As we noted in connection with a similar example in chapter 3, one might, of course, replace a particular raven (contained, say, in a cage in my living room) with a differently colored bird, for example, a blue jay. But, intuitively, this does not amount to changing the raven into a blue jay or manipulating its color. According to the manipulationist account of explanation I am proposing, it is only when one has identified conditions relevant to the manipulation of a raven's color that one has provided an explanation of why it is black.

(5.1.1) is thus, on the manipulationist theory, a defective explanation, or at least lacks a crucial feature that a satisfactory explanation ought to possess. I say more to support this judgment below. For now, I simply want to point out that the manipulationist account contrasts interestingly with *DN* accounts of explanation. According to my account, if a putative explanation cites a condition that is nomologically sufficient for an outcome, but there is no well-defined

notion of changing that condition or the explanation does not correctly describe how the outcome would change under changes in the condition, the explanation is unsatisfactory. By contrast, standard versions of the *DN* model will regard the citing of such a sufficient condition along with an appropriate covering generalization as a satisfactory explanation. As we shall see, the demand that a successful explanation answer *w-questions* is tantamount to the requirement that explanations must provide information about the causes of their explananda, if “cause” is understood along the manipulationist lines described in chapter 2. Thus, another way of putting the contrast that I have sought to draw is that providing a nomologically sufficient condition for an outcome is different from providing a cause of that outcome, and it is the latter that is crucial to explanation.

## 5.2 Minimal Covering-Law Explanation

Consider another example that further illustrates the contrast between (5.1.1)–(5.1.2) and (5.1.3)–(5.1.4) that I am trying to capture. In the following passage in “Aspects of Scientific Explanation,” Hempel (1965a) describes two different explanations, which I have distinguished by means of brackets for later reference:

A purely logical point should be noted here, however. If an explanation is of the form (*D-N*), then the laws  $L_1, L_2, \dots, L_r$  invoked in its explanans logically imply a law  $L^*$  which by itself would suffice to explain the explanandum event by reference to the particular conditions noted in the sentences  $C_1, C_2, \dots, C_k$ . This law  $L^*$  is to the effect that whenever conditions of the kind described in the sentences  $C_1, C_2, \dots$  the explanandum-sentences occur. Consider an example: A chunk of ice floats in a large beaker of water at room temperature. Since the ice extends above the surface, one might expect the water level to rise as the ice melts; actually, it remains unchanged. Briefly, this can be explained as follows: [According to Archimedes’ principle, a solid body floating in a liquid displaces a volume of liquid that has the same weight as the water displaced by its submerged portion. Since the melting does not change the weight, the ice turns into a mass of water of the same weight, and hence also of the same volume, as the water initially displaced by its submerged portion; consequently, the water level remains unchanged. The laws on which this account is based include Archimedes’ principle, a law concerning the melting of ice at room temperature; the principle of the conservation of mass, and so on.] None of these laws mentions the particular glass of water or the particular piece of ice with which the explanation is concerned. Hence the laws imply not only that as this particular piece of ice melts in this particular glass, the water level remains unchanged, but rather the general statement  $L^*$  that [under the same *kind* of circumstance, i.e., when any piece of ice floats in water in any glass at room temperature, the same *kind* of phenomenon will occur, i.e., the water level will remain unchanged] . . . clearly,  $L^*$  in conjunction

with  $C_1, C_2, \dots, C_k$  logically implies  $E$  and could indeed be used to explain, in this context, the event described by  $E$ . (p. 347)

The first set of brackets encloses a sketch of what might naturally be described as a theoretical or scientific explanation of  $E$ . The second encloses what Hempel calls a minimal covering-law explanation of  $E$ . The contrast between these two explanations is, I think, quite striking. The first seems to provide a genuine understanding of why  $E$  occurs, whereas the second provides little if any understanding. In part, my discussion in this chapter is an attempt to explicate this contrast.

Note to begin with the contrast between  $L^*$  and the laws—Archimedes' principle, the law of conservation of matter, and so forth—which figure in the explanation enclosed in the first set of brackets. These laws are such that they can be used in conjunction with different sets of initial conditions to answer a range of what-if-things-had-been-different questions about different explananda; thus, Archimedes' principle can be used, for example, to explain why any solid body floating in a liquid (not just ice floating on water) displaces the volume of liquid it does. Because of this, the explanation enclosed in the first set of brackets gives us some sense for the range of conditions under which the explanandum will hold. When we see the relevance of the considerations invoked in the explanans—that a solid that floats displaces a volume of liquid equal to the weight of its submerged portion, that no change of weight is involved when a solid melts—we see that the same reasoning could well be used to show that any solid floating in its liquid form and melting will leave the level of the liquid unchanged. Nothing in the explanation of  $E$  turns essentially on the fact that a piece of ice floating in water is involved, nor does it matter that the surrounding liquid is at "room temperature" rather than at any other temperature that is consistent with the ice melting, or that the system is in a "glass" rather than some other container. We can also see, once we appreciate the relevance of Archimedes' principle to this case, that it does make a difference whether the solid is floating in the liquid. If, for example, the solid sinks in the liquid, then it will displace a volume of liquid equal to its own volume rather than its weight, and because, for almost all solids, a given mass will increase in volume when it melts, the level of liquid in the container will rise as the submerged solid melts in such cases.

By contrast, the minimal covering-law explanation in terms of  $L^*$  gives us no information of this kind. It can be regarded as assuring us, perhaps, that the water level in the glass "had" to remain unchanged when the ice melted, but it does not answer the *w-questions* that the explanation enclosed in the first set of brackets can be used to answer. And to say this is to say that the minimal covering-law explanation does not perspicuously exhibit the factors or conditions that are relevant to the water level in the glass remaining unchanged in the way that the explanation enclosed in the first set of brackets does. One can be aware of the information contained in the minimal covering-law explanation, and yet largely fail to understand what it is about an ice cube melting in a glass that results in the water level in the glass remaining unchanged.

One may be aware of  $L^*$  and yet still be unclear whether the fact that the water level in the glass remained unchanged has to do with some feature unique to water and ice, whether it has to do with the fact that a glass rather than some other container was employed, whether it has to do with whether the system is at room temperature, and so forth. Indeed, when taken at face value, the minimal covering-law explanation wrongly or misleadingly suggests that this information is relevant to the phenomenon we are trying to explain when in fact it is not. The contrast between these two explanations is precisely the contrast between an explanation that satisfies the conditions for successful *DN* explanation (it exhibits a nomologically sufficient condition for its explanandum) but fails to accurately answer a relevant range of what-if-things-had-been-different questions (or at least an appropriately “large” range of such questions; see below) and an explanation that does answer such questions.

### 5.3 Explanation and Counterfactuals

When we ask what would happen to the strength of an electromagnetic field or to the behavior of a firm under a range of initial conditions besides those that actually obtain in (5.1.3) and (5.1.4), we are plainly concerned with the truth of various counterfactuals. The theory of causal explanation I have been sketching is thus one that ties explanatory import very closely to the provision of certain kinds of counterfactual information: it might fairly be described as a counterfactual theory of causal explanation. In view of the discussion in chapters 2 and 3, readers will not be surprised to learn that I take the relevant counterfactuals to be counterfactuals that describe the outcomes of interventions. For example, the counterfactuals that matter for the explanatory import of (5.1.3) concern what would happen to the field if we (or some natural process) were to physically intervene in the system in question: increasing the charge density on the wire by connecting it to an appropriate source, or changing its geometry by twisting it into a circle or a solenoid. Coulomb’s law does not just describe a correlation or association between quantities like charge density, distance, and field strength; instead, it tells us how we (or nature) can change the field strength by altering  $\lambda$  and  $d$ . Thus, (5.1.3) conveys information that is relevant to manipulating or controlling the field. Parallel remarks apply to (5.1.4). The above account of how (5.1.3)–(5.1.4) work and why they have explanatory import is thus meant to illustrate the connections between causal explanation, manipulation, and control emphasized in previous chapters.

This point that the counterfactuals that matter for successful causal explanation are counterfactuals that have an interventionist interpretation dovetails with the observation, made by both Nancy Cartwright (1983b) and Paul Humphreys (1989), that not every logical or mathematical transformation one can perform on a formal representation of a system corresponds to a physical manipulation performed on the system itself. The mere fact that our theory of some system is such that, from the assumption that  $P$  is true, we can derive

that  $Q$  is true, does not mean that physically intervening to change whether  $P$  is true will change whether  $Q$  holds, even if the derivation makes essential use of a law.

Two of the classic counterexamples to the *DN* model discussed in chapter 4 provide illustrations of this point and also provide additional motivation for the theory of explanation sketched above. Consider first the problem of explanatory asymmetries. As noted, there is general agreement that one may appeal to the relationship

$$(5.3.1) \quad T = 2\pi\sqrt{\ell/g}$$

to explain the period of a simple pendulum in terms of the length. However, one may also invert (5.3.1) to obtain  $l$  as a function of  $T$  and  $g$ :

$$(5.3.2) \quad \ell = \frac{T^2 \cdot g}{4\pi^2}$$

Given the counterfactual supposition that the period  $T$  of pendulum  $P$  has increased to  $T^*$ , with the gravitational field  $g$  remaining unchanged, one may use (5.3.2) to derive what the new length  $\ell^*$  would have to be to account for the increase in period. There is thus *some* interpretation of the counterfactual

$$(5.3.3) \quad \text{If the period of } P \text{ were } T^*, \text{ then the length would be } \ell^*$$

such that (5.3.3) comes out true. Nonetheless, one may not appeal to the period to explain the length.

The account that I favor traces this explanatory asymmetry to an underlying physical asymmetry in the roles played by the length and the period, an asymmetry connected to facts about manipulation and control. The asymmetry is this: suppose that one physically intervenes to change the length with respect to the period (by, e.g., stretching the wire connected to the weight on the end of the pendulum). Recall that to qualify as an intervention on  $\ell^*$  with respect to  $T$ , this change must be causally independent of and uncorrelated with  $g$ , the other cause of  $T$ . This condition will be met by any ordinary stretching (increase in the length) of the wire that leaves the gravitational field  $g$  unchanged. Under such an intervention, the period of the pendulum will change. It is thus true, in the relevant sense, that the period is counterfactually dependent on the length and that we may appeal to the change in length to answer a range of questions about the conditions under which the period would have been different. This vindicates the judgment that the length of the pendulum figures in an explanation of its period. By contrast, there are no interventions on the period  $T$  that will change the length of the pendulum. It is true that one might manipulate  $T$  by moving the pendulum to a different gravitational field  $g^*$  (without changing its length in any other way except via the effects, if any, of this change in location), but this will not result in a change in the length of the pendulum. Of course, one might also change the period of the pendulum by changing its length, but it should be clear from the

characterization in chapter 3 that this is not an intervention on the period with respect to the length. In other words, any manipulation (e.g., moving the pendulum to a different gravitational field) that changes the period via a route on which the length is not an intermediate variable between the intervention and the period will not change the length, whereas any process that changes the period will do so via a route on which the length is an intermediate variable and hence will not count as an intervention on the period with respect to the length. Because there are no interventions on the period that will change the length, one cannot appeal to the period to explain the length. This difference is reflected in the commonsense judgment that, although one can change the period by changing the length, one cannot change the length by changing the period. This is a real asymmetry which we represent by writing (5.3.1) rather than (5.3.2).

We may think of this example as again illustrating the general point, reflected in a different way by (5.1.1)–(5.1.2), that providing a nomologically sufficient condition for an outcome is not the same thing as answering a set of *w-questions* about that outcome when this latter notion is interpreted along interventionist lines. The derivation of the length of the pendulum from its period and (5.3.2) provides a nomologically sufficient condition for  $\ell$  but does not (in the relevant sense that has to do with what would happen under interventions) identify conditions such that changing these will change the length.

Essentially the same diagnosis holds for the problem of explanatory irrelevancies. Consider again

(5.3.3) L 5.3.3 All men who take birth control pills regularly fail to get pregnant.

C 5.3.3 Mr. Jones is a man who takes birth control pills regularly.

M 5.3.3 Mr. Jones fails to get pregnant.

There are two possible ways of interpreting (L 5.3.3). First, if taken literally, (L 5.3.3) says only that male pill takers fail to get pregnant. It says nothing about what would happen if males were not to take pills, or about the conditions under which an alternative to failure to get pregnant (i.e., pregnancy) would occur. So interpreted, (5.3.3) does not even purport to tell us about the conditions under which (M 5.3.3) would have been different. An alternative interpretation of (L 5.3.3) is that it suggests (even though it does not literally say) that changing whether a male takes birth control pills will change whether he gets pregnant, a claim that is, of course, false. Moreover, the other condition cited in the explanans of (5.3.3) (being a man) is, arguably, like being raven: as argued in chapter 3, it isn't clear (at least without additional specification) what is meant by changing this condition and, to the extent that this is so, counterfactuals about what would happen under such a change will lack determinate truth values. Thus, in either interpretation, (5.3.3) fails to convey correct information about the conditions under which its explanandum would have been different. On the second interpretation, it also

misleadingly suggests that a condition cited in its explanans (taking birth control pills) is a condition such that changes in it would lead to changes in whether Jones gets pregnant, when in fact this is not true. This failure is reflected in our judgment that taking birth control pills is explanatorily irrelevant to whether Jones gets pregnant.

Now contrast

(5.3.4) L 5.3.4 All women who meet condition *K* (*K* has to do with whether the woman is fertile, has been having intercourse regularly, and so forth) and who take birth control pills regularly will not get pregnant, and furthermore all women who meet condition *K* and do not take birth control pills regularly will get pregnant.

C 5.3.4 Ms. Jones is a woman who meets condition *K* and has been taking birth control pills.

M 5.3.4 Ms. Jones does not get pregnant.

Here we have considerably more inclination to say that at least a crude explanation of (M 5.3.4) has been provided. I take the difference between (5.3.3) and (5.3.4) to be a reflection of the fact that (5.3.4) shows us, whereas (5.3.3) does not, how, if conditions had been different, a different outcome would have ensued. That is to say, the conditions cited in the explanans of (5.3.4) (whether or not *K* obtains, whether or not the subject takes birth control pills regularly) are such that changes in them would lead to changes in the outcome being explained: if Ms. Jones stops taking birth control pills, is fertile, and has intercourse, she will or at least may get pregnant; if she fails to take the pills but also doesn't have intercourse, she will not get pregnant; and so on. (L 5.3.4) draws our attention to this systematic pattern of dependency of changes in the explanandum of (5.3.4) on changes in its explanans: we can combine (L 5.3.4) with one set of claims about initial conditions to deduce that Ms. Jones will get pregnant under those conditions and with a different set of initial conditions to deduce that Ms. Jones will not get pregnant under those different conditions. Unlike (5.3.3), (5.3.4) thus locates its explanandum within a range of possible alternatives and shows, at least in a crude way, the range of conditions under which this explanandum would hold and what sorts of changes in those conditions would instead lead to one of these alternatives. In contrast to (5.3.3), (5.3.4) thus shows us why a certain condition being what it is, the explanandum of (5.3.4) rather than certain alternatives was realized, and in doing so shows how this condition (Ms. Jones's taking birth control pills) makes a difference for, or is relevant to, the explanandum. This feature of (5.3.4), a feature that it shares with (5.1.3)–(5.1.4) but not with (5.1.1)–(5.1.2), is, I have been arguing, central to our sense that it identifies conditions that are explanatorily relevant to its explanandum, whereas (5.3.3) does not. The difference between (5.3.3) and (5.3.4) also further illustrates the connection between causal explanation and information potentially relevant to manipulation and control: (5.3.4) does, whereas (5.3.3) does not, provide information (or identify conditions) that are potentially relevant to controlling whether or not Jones becomes pregnant.



In a discussion of examples like (5.3.3), Wesley Salmon (1984, 1989) has drawn attention to the important general point that although the addition of irrelevant information to a sound deductive argument does not undermine the soundness of the argument, the addition of such information can be fatal to (or at least can seriously undermine) the goodness of an explanation, as (5.3.3) illustrates. The theory of causal explanation sketched above provides a natural analysis of what an explanatory irrelevancy involves in simple structures of the sort under discussion: it is a condition such that there are no changes in it (due to interventions) that are associated with corresponding changes in the explanandum phenomenon. Put slightly differently, explanatory irrelevance is understood counterfactually: in examples of the sort under discussion, an explanans variable  $S$  is explanatorily irrelevant to the value of an explanandum variable  $M$  if  $M$  would have this value for any value of  $S$  produced by an intervention. Given the theory of explanation sketched above, there is no mystery about why such irrelevancies undermine explanatory import or about why, as we saw in chapter 4, providing information that is explanatorily relevant is a different matter from providing nomic grounds for expecting. More generally, we may say that the counterfactual account of explanation sketched above provides a natural (and unified) diagnosis of what is wrong with putative explanations like (5.3.3) that contain explanatory irrelevancies and with explanations like (5.3.2) that fail to respect explanatory asymmetries; in both cases, their inadequacy as explanations may be traced to their failure to answer any *w-questions*.

The characterization of explanatory irrelevancies given in the preceding paragraph works only for relatively simple examples like (5.3.3). In more complex structures with multiple causal connections, a more complex characterization of the notion of an explanatory irrelevancy will be required. Consider the example from chapter 2, section 3, in which there is exact cancellation along two different routes of the effects of ingestion of birth control pills  $B$  on thrombosis  $T$ . Here interventions on the value of  $B$  will not change the value of  $T$ , but  $B$  is explanatorily relevant to  $T$ . Following the strategy of chapter 2, we can characterize the explanatory relevance of  $B$  to  $T$  in terms of the idea that there is at least one intervention that in changing the value of  $B$  will change the value of  $T$  when other (off-path) variables, in this case  $P$ , are fixed by interventions at some value. More generally, explanatory relevance in complex structures with multiple connections can still be characterized in terms of facts about patterns of counterfactual dependence, but the counterfactuals in question will need to make reference to what will happen under combinations of interventions, along the lines described previously.

Salmon (1984) goes on to conclude, from the fact that irrelevancies undermine explanations, but not arguments, that no explanations are arguments. In his more recent survey (1989), he retreats from this conclusion: "In my zeal to rebut the claim that all explanations are arguments, I argued that no explanations are arguments. This view now seems too extreme; as it seems to me now, some are and some are not" (p. 101). I fully agree with this remark, but would offer a gloss that is probably different from what Salmon intended. As I

see it, the underlying or unifying idea in the notion of causal explanation is the idea that an explanation must answer a what-if-things-had-been-different question, or exhibit information about a pattern of dependency. One way, but by no means the only way, of doing this is by constructing a derivation (or series of derivations) or a deductive argument, in the manner of (5.1.3)–(5.1.4). In such cases, one shows how changes in an explanandum phenomenon (e.g., the field intensity) changes with changes in the conditions cited in the explanans conditions (the geometry of the conductor, the charge density, etc.) by deriving or deducing different explananda (different expressions for the field) from different sets of initial and boundary conditions using the same generalization (Coulomb's law). The idea is that these derivations trace or mirror the relations of physical dependency that hold between the explanans conditions and the explananda phenomena—relations that would be revealed if, for example, we were to physically intervene to alter the explanans conditions. However, such explicit derivations are not the only way an explanation can answer a *w-question* or convey information about a pattern of dependency. As we will see in more detail below, singular-causal explanations also convey such information, but (at least often) not by exhibiting derivations. Other representational devices, such as diagrams and the directed graphs employed in chapters 2 and 3, can also convey information about dependency relations, even though they do not take the form of explicit deductive arguments. The account of explanation sketched above thus supports the idea (which I take to be hard to deny) that deductive or derivational structure plays an important role in some causal explanations, without imposing the implausible requirement that all causal explanations must take the form of deductive (or, for that matter, inductive) arguments.

Thus, although I agree with proponents of the *DN* model that deductive structure makes an important contribution to explanatory import in some explanations, it also should be clear that my view of the role of derivational structure in explanations like (5.1.4) is quite different from the role emphasized in the *DN* model. According to the *DN* model, derivational structure matters because it involves the exhibition of a nomologically sufficient condition for the explanandum phenomenon or, more broadly, because it involves the exhibition of nomic grounds for expecting. On the contrasting view I wish to defend, derivational structure is relevant to causal explanation because it is used to show how an explanandum phenomenon would have been different if the conditions cited in an explanans had been different in various ways. That this is different from providing nomic grounds for expecting is established by the fact that (5.1.1) and (5.1.2) as well as (5.3.2) and (5.3.3) all provide the latter kind of information but not the former. It is also established by the fact that, as discussed below, there are other forms of explanation (such as singular-causal explanations or explanations that appeal to generalizations that are not laws) that can answer *w-questions* and trace dependency relations even when they fail to provide nomologically sufficient conditions.

Before leaving the topic of the role of derivational structure in explanation, let me underscore in a slightly different way a point made above. Although

derivational structure sometimes plays an important role in explanation, it is *not* true that whenever different initial conditions can be combined with the same generalization to deduce a range of different explananda, one has an explanation of those explananda; the derivation (5.3.2) of the length of a pendulum from its period represents an obvious counterexample to this claim. Instead, what matters for purposes of causal explanation is what the real dependency relations in the world actually are: what would in fact happen to some potential explanandum if other conditions were changed in various ways. Invoking such independent existing dependency relations commits us to the kind of modest realism about causal and explanatory relationships described in chapter 3, but to nothing more elaborate. This sort of realism commits us to claims like the following: in the physical situation described by (5.1.3) there is a determinate fact about what would happen to the field if we were to increase the charge density along the wire or twist it into a circle; our explanation of the field intensity must capture or represent such facts if it is to be a good one. Not all derivations from true premises will trace or mirror dependency relations, as the derivation of the length of a pendulum from its period illustrates. The theory I am proposing is thus what Salmon (1984) calls a realist or “ontic” theory of explanation, rather than an epistemic or logic-based model, in the sense that it insists that one cannot read off just from the logical or deductive relations between a putative explanans and explanandum whether the former genuinely explains the latter. It is physical dependency relations, as expressed by the relevant counterfactuals about what would happen under interventions, that are primary or fundamental in causal explanation; derivational relations do not have a role to play in explanation that is independent or prior to such dependency relations, but rather matter only insofar as (or to the extent that) they correctly represent such relationships. I return to this point in the discussion of unificationist accounts of explanation in chapter 8.

Drawing together the various strands of our discussion and confining our attention to simple structures without multiple connections, we may summarize the account of explanation sketched above as follows. As with the *DN* model, an explanation relates two components, an explanans and an explanandum; in particular, it exhibits how the latter is counterfactually dependent on the former. Both explanans and explanandum are formulated in terms of variables. The explanandum is a true (or approximately true) proposition to the effect that an explanandum variable takes on some particular value. We may think of the explanans as a set of propositions, some of which specify the actual (approximate) values of explanans variables, and others of which specify relationships between the explanans and explanandum variables. These relationships must be change-relating: they must relate changes in variables in the explanans to changes in the explanandum variable, and in particular, they must correctly describe how the explanandum variable would change under interventions on the explanans variables. More specifically, there must be some intervention that in changing one or more of the explanans variables from their actual value changes the value of the explanans variable from its actual value. The idea is thus this:

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

I emphasize that this is a minimal condition for successful explanation; ideally, a successful explanation will involve a generalization  $G$  and explanans variable(s)  $X$  such that  $G$  correctly describes how the value of  $Y$  would change under interventions that produce a range of different values of  $X$  in different background circumstances.

#### 5.4 Subsumption versus Dependency

Consider again the unflattering contrast drawn between (5.1.1)–(5.1.2) and (5.1.3)–(5.1.4). The suggestion that (5.1.1) is not a satisfactory explanation of why some particular raven is black, or at least that (5.1.1) lacks a crucial feature that is relevant to explanatory import, may seem deeply counter-intuitive to those whose intuitions have been nourished on the *DN* (or related) models of explanation. Many readers may initially respond that to explain an outcome just *is* to fit it into a more general pattern or to subsume it under a regularity. (5.1.1) plainly accomplishes this and so, contrary to what I have suggested, it must be an explanation. Let me therefore try to say more to motivate the idea that (5.1.1) is explanatorily unsatisfactory, or at least qua explanation omits something of crucial importance.

Consider someone who asks for an explanation of why some raven is black. Underlying the conception of explanation defended above is the idea (contrary to what the *DN* model suggests) that such a person (let us call him  $Q$ ) does not just wish for a demonstration via a law and some statement of initial conditions that this raven “had” to be black. Rather, when  $Q$  asks for an explanation of why some raven is black, he wishes to know what it is about that raven that makes it black.  $Q$  is puzzled because he is unable to identify those features of a raven on which its pigmentation depends and is unaware of the laws or generalizations that describe this dependency. When  $Q$  is in such a situation, he will not be helped by being told that all ravens are black. This generalization tells  $Q$  that all other ravens have the feature he finds puzzling about this particular raven, but it does not tell him what that feature depends on, which is what he wishes to know.

We can bring this out more clearly by contrasting (5.1.1) with a more satisfying explanation of why some particular raven is black, the sort of

explanation we might naturally describe as scientific or theoretical. In my view, such an explanation would involve something like this: an identification of those specific biochemical reactions within ravens that produce their distinctive pigmentation, and a specification of the genetic mechanisms that are responsible for those reactions. An explanation of this kind (let us call it 5.4.1) would exhibit the structural features to which I called attention above. That is, it would locate its explanation within a range of alternatives and would answer a range of *w-questions* about the conditions under which the explanandum and these alternatives would have ensued. Put differently, such an explanation would identify mechanisms that could be used to explain why a nonblack raven (or raven-like bird) has the color it has. That is, it would identify features of the raven genotype and of the biochemical pathways involved in the synthesis of pigmentation, such that changes in these would lead to ravens that were not black in color. It might also be an account that could be used to explain why birds of other, related species have the colors they do. It would, in any event, be an account that identified the range of conditions under which the present typical coloration of ravens would continue to hold, and that would make it clear how, if the genetic structure of ravens or their biochemistry were to alter in certain ways, their color would also change. Unlike the information conveyed by (5.1.1), this is information that is potentially relevant to the control or manipulation of a bird's color. I take it to be uncontroversial that such an explanation strikes most of us as deeper and more illuminating than (5.1.1). It seems to me to be an important advantage of the account of explanation described above that it captures or draws attention to these further features that a satisfying explanation of why some particular raven is black ought to possess, and in doing so, also draws our attention to an inadequacy in (5.1.1).<sup>1</sup>

## 5.5 Additional Motivation for the Manipulationist Conception

Although I hope to have said enough to make the idea that explanations must provide what-if-things-had-been-different information intuitively appealing, it is natural to wonder whether there is anything more systematic that can be said in support of this idea. One motivation is implicit in the discussion of the preceding sections. Intuitively, to causally explain a phenomenon is to provide information about the factors on which it depends and to exhibit how it depends on those factors. This is exactly what the provision of counterfactual information of the sort described above accomplishes: we see what factors some explanandum *M* depends on (and how it depends on those factors) when we have identified one or more variables *S* such that changes in these (when produced by interventions) are associated with changes in *M*. For example, we see what the intensity of the field in (5.1.3) depends on when we see how changing the charge density, the geometry of the conductor, and so on will change the field intensity. Similarly, the fact that whether or not Mr. Jones becomes pregnant does not depend on whether he takes birth control pills

(and hence, that this factor is not part of a causal explanation for (M 5.3.3)) is reflected in the fact that changing this factor will not change whether Jones becomes pregnant.

A second, closely related motivation is rooted in the observation that it is natural to think that successful explanation often has to do with the exhibition of causal relationships, broadly construed. If, as argued in chapters 2 and 3, the content of causal claims is closely tied to the truth of various counterfactuals, then it also seems plausible that, insofar as explanation is causal, it will involve the exhibition of patterns of counterfactual dependence. The version of the manipulability theory developed in chapters 2 and 3 captures what I intend by the notion of a causal relationship "broadly construed," and also shows why the notions of causation and explanation are so closely intertwined: causal claims are explanatory in virtue of providing the sort of counterfactual information that is at the heart of successful explanation. We can also see, as suggested in chapter 4, that the common diagnosis that the problems posed by the existence of explanatory asymmetries and explanatory irrelevancies stem from the failure of models like the *DN* model to incorporate causal information is correct, if causal information is understood along the lines sketched in chapters 2 and 3. Thus, on my view, it is perfectly true that the reason why citing Jones's failure to take birth control pills does not explain his failure to get pregnant is that the former does not cause the latter, as long as the notion of cause is understood along manipulationist lines and not in some other way. Similarly for the failure of the period of a simple pendulum to explain its length.

As noted in chapters 2 and 3, the paradigmatic causal relationship most often discussed by philosophers involves dichotomous variables, as in the relationship between whether or not a diseased patient is treated with some drug and whether or not he recovers. Suppose for the sake of argument that this relationship, call it *R*, is deterministic: patients who are treated always recover (including in those cases in which the treatment is imposed by an intervention), and those who are not treated never recover. Then, according to the account of explanation sketched above, we may explain why some particular patient *A* recovers by citing *R* and the fact that he has been treated and why some other patient *B* has not recovered by citing *R* and the fact that *B* has not been treated. In both cases, a *w-question* about the explanandum phenomenon has been correctly answered: if *A* had not been treated, he would not have recovered, and if *B* had been treated, he would have recovered. If instead the relationship *R* is indeterministic, then it is less transparent how the what-if-things-had-been-different condition should be understood (see section 5.8), but it is perhaps uncontroversial that we can at least appeal to *R* to explain why the expected rate of recovery in the treatment group is higher than that in the control group. Here too, this explanation is explanatory in virtue of conveying information about the conditions under which this explanandum would have been different: the expected rate of recovery would not have been higher in the treatment group if treatment had been withheld from this group. Finally, note that this example provides an additional illustration of why it is not enough for

treatment to explain recovery that treatment be nomologically sufficient for recovery or that recovery be nomically expectable under treatment with high probability. These conditions would be satisfied if everyone spontaneously recovers from the disease regardless of whether they are treated. What is required instead if treatment is to figure in an explanation of recovery is that treatment be relevant to or make a difference to recovery, and this notion is cashed out in terms of the what-if-things-had-been-different requirement.

We may think of the two examples (5.1.3) and (5.1.4) of theoretical explanation with which we began this chapter as involving a generalization or extension of the pattern dependence ( $R$ ) between treatment and recovery. Whereas the variables that figure in ( $R$ ) are two-valued, the variables in (5.1.3) and (5.1.4) can take any real value in certain ranges. Whereas ( $R$ ) shows us how one possible change (from treatment to nontreatment or vice versa) is associated with a change in whether recovery occurs (or in the probability of recovery), (5.1.3) and (5.1.4) show us instead how any one of a great number of changes in their explanans variables will lead to one of many possible changes in their explanandum variables. In other words, (5.1.3) and (5.1.4) give us information about a much more detailed and fine-grained quantitative pattern of counterfactual dependence than the “binary” pattern described by ( $R$ ). According to the analysis in chapter 2, such information is a form of nonbinary causal information which has the virtue of explicitly spelling out, in precise quantitative terms, exactly how various interventions that change the values of various explanans variables will change the values of explanandum variables.

## 5.6 The Role of Laws

I noted above that the generalization (Coulomb's law) that figures in (5.1.3) is often regarded as a law of nature. The generalization about profit maximization that figures in (5.1.4) is less commonly regarded as a law, but it nonetheless has a kind of generality as well as other features that philosophers have associated with laws. (For example, it plays a fundamental unifying role in economic theory.) However, as noted in chapter 4, there are many explanations that do not seem to “involve” laws at all, assuming that we have even a modestly demanding conception of what a law is, and much of the appeal of the account of explanation sketched above derives from the fact that it can be used to provide a natural and satisfying treatment of such explanations. This claim is defended at greater length in chapter 6, but it will be useful to introduce some simple illustrations at this point.

Consider again the generalization ( $R$ ) “Treatment with  $D$  causes recovery.” As argued in the previous section, one may appeal to ( $R$ ) to explain the higher expected incidence of recovery in a particular treatment group that receives  $D$  in comparison with a control group from whom  $D$  has been withheld. Nonetheless, ( $R$ ) lacks many of the features commonly associated with laws of nature. Although we may be satisfied that it holds in the particular experimental

context under discussion (internal validity), we may be completely unable to delimit the conditions under which (R) holds outside this experimental context. In addition, (R) is vague and imprecise. Although a number of philosophers have argued that (R) nonetheless is either a law or serves as a surrogate or stand-in for a law, my view is that such arguments do little to illuminate the role that (R) plays in explanation. As long we can appeal to (R) to answer a set of *w-questions*, we don't need to assume the burden of arguing that (R) is a law of nature to vindicate its explanatory status. There would be a serious motivation for regarding generalizations like (R) as laws only if there were some reason for assuming that only laws can figure in answers to *w-questions* and, as we shall see in chapter 6, there is no reason to accept this assumption.

Consider another example. Eric Veblen (1975; discussed at length in Achen 1982) investigates the effect of editorial endorsements by a particular newspaper, the *Manchester Union Leader*, on New Hampshire elections during the period 1960–1972. He regresses a variable *V* (vote difference) measuring the difference between the vote for the *Union Leader* candidate in Manchester (where the *Union Leader's* circulation is large) and the vote for this candidate outside of the Manchester area (where the newspaper's circulation is low) against a dichotomous variable *S* (slant), which takes the value 0 or 1 depending on whether the bias in favorable coverage is above or below average:

$$(5.6.1) \quad V = b_0 + b_1S + U$$

Veblen finds that  $b_0 = -11$  percent and  $b_1 = 22$  percent. That is, the estimated effect of favorable coverage is quite large: a change from below- to above-average slant is associated with a 22 percent increase in the vote for the *Union Leader's* candidate.

Both Veblen and Achen regard this relationship as causal, and claim that one can explain facts about the vote that various candidates receive by appealing to facts about the *Union Leader's* editorial policies. We will explore a number of issues concerning the interpretation of equations like (5.6.1) and the sort of evidence that is required to support them in chapter 7. Here, I want only to suggest that for Veblen's claim about the explanatory status of (5.6.1) to be correct, all that is required is that, given certain background conditions characteristic of New Hampshire during the period 1960–1972, (5.6.1) correctly (or approximately correctly) describes a pattern of counterfactual dependence linking interventions involving changes in editorial policies from above- to below-average slant to the vote difference a candidate receives. When this is true, one may appeal to (5.6.1) to answer a range of what-if-things-had-been-different questions about the vote a candidate receives. For example, one has good grounds for believing that if a candidate favored by the *Union Leader* had not received such favorable treatment, his vote difference would have been considerably smaller. In other words, my claim is that one may regard an equation like (5.6.1) linking slant to vote as explanatory in



virtue of (and to the extent that it is true that) it exhibits the same sort of pattern of systematic counterfactual dependence we found in (5.1.3) and (5.1.4). I also suggest that (5.6.1), like  $(R)$ , is not a very plausible candidate for a law of nature. I will say more to support this assessment in chapter 6, but for present purposes we may note that the scope of (5.6.1) is very narrow and there are good reasons to doubt that it describes a relationship between newspaper bias and votes that would generalize to other places and times. Moreover, it is clear that even in New Hampshire during the period 1960–1972, there are many possible occurrences that would have disrupted (5.6.1). Nonetheless, as I suggest in chapter 7, we can have good reasons for believing that (5.6.1) can be used to correctly answer a set of *w-questions*—and hence that it is explanatory—despite lacking the standard features of a law of nature. Again, the account of explanation sketched above better captures how explanations that appeal to generalizations like (5.6.1) work than nomothetic models.

## 5.7 Explanations as Change-Relating

I noted above that it is built into the manipulationist account of explanation I have been defending that explanatory relationships must be change-relating: they must tell us how changes in some quantity or magnitude would change under changes in some other quantity. Thus, if there are generalizations that are laws but that are not change-relating, they cannot figure in explanations. This represents another respect in which the manipulationist account differs from the *DN* model, which imposes no requirement that the generalizations that figure in explanations must be change-relating. As an illustration, consider the following generalization:

(5.7.1) All physical processes propagate at a speed less than or equal to that of light.

(5.7.1) is an uncontroversial example of a law. Suppose that a particle is expelled from a star during a supernova. Despite the enormous energy that is unleashed during this explosion, we still find the particle receding from the former location of the star at a speed that is less than the speed of light. (For definiteness, let us assume that we measure velocity relative to the frame in which the star was at rest.) Why is this so? Is it not an explanation to cite the fact that the moving particle constitutes a genuine physical process, and that all physical processes propagate at luminal or subluminal velocities?

According to the manipulationist account, this is no explanation. To see why, we must identify the relevant variables and potential interventions on those variables. Insofar as the notion of a physical process contrasts with anything at all, it presumably contrasts with the notion of a pseudo-process (such as a shadow or a spot of light on a wall). The explanans variable will thus have the values  $\{\textit{physical process}, \textit{pseudo-process}\}$ . An intervention would then have to be something that causes the particle in question to be transformed

from a physical process to a pseudo-process, or from a pseudo-process to a physical process. As suggested in chapter 3, we have no clear conception of what such an intervention would involve. Moreover, even if such an intervention were possible, (5.7.1) tells us nothing about the conditions under which an alternative to luminal or subluminal (i.e., superluminal) propagation would occur. It follows that purported explanations employing (5.7.1) fail to answer any what-if-things-had-been-different questions about the (sub)luminal velocity of the particle. In an obvious sense, explanations that appeal to (5.7.1) fail to identify the conditions on which the (sub)luminal velocity of the particle counterfactually depends. To the extent that saying what an outcome depends on is at the heart of successful explanation, (5.7.1) fails to be explanatory. This example provides an additional illustration of the great difference between providing a nomologically sufficient condition for an outcome and specifying what that outcome depends on.

The judgment that (5.7.1) is unexplanatory may seem counterintuitive to some. By way of mitigation, note that a genuinely explanatory generalization may be found in the neighborhood of (5.7.1). Consider the generalization relating the kinetic energy imparted to the particle receding from the supernova (relative to some specific frame of reference, such as the initial rest frame) and its velocity, according to which the velocity of the particle approaches the speed of light asymptotically as kinetic energy approaches infinity. This generalization, I contend, *is* genuinely explanatory and could be used to explain the actual subluminal velocity possessed by the particle. It *can* be used to answer what-if-things-had-been-different questions: it tells us how changes in the velocity of the particle will counterfactually change, according to the kinetic energy it receives. So it is not that the actual subluminal velocity possessed by the particle is inexplicable; it just isn't explained by (5.7.1).

## 5.8 Singular-Causal Explanation

It will be recalled that chapter 4 argued that an attractive feature of the *DN/IS* model was that it attempted to exhibit a continuity or similarity between more “scientific” or “theoretical” explanations like (5.1.3)–(5.1.4) and other sorts of explanations like singular-causal explanations or the explanation conveyed by the regression equation (5.6.1) relating newspaper bias to vote. However, we found the particular analysis associated with the *DN* model of what that continuity consisted in—namely, that singular-causal explanations explain in virtue of conveying information about underlying *DN* or *IS* (or for that matter nomothetic) structures—to be unpersuasive. The account of explanation sketched above may be regarded as a quite different proposal regarding what that continuity between (5.1.3) and explanations like (5.6.1) consists in, one that identifies the common element with the exhibition of the right sort of pattern of counterfactual dependence between explanans and explanandum variables. In this section, I briefly extend this account to singular-causal (or token-causal) explanations, showing how these also fit into the general

framework sketched above. Doing so will also illustrate how the manipulationist account can be applied to explanations that do not have the structure of deductive arguments.<sup>2</sup>

In the special case in which singular-causal claims are evaluated with respect to a known background deterministic structural model, as with the examples discussed in section 2.7, it should be obvious how the manipulationist account applies. Token-causal claims imply various counterfactuals, and it is in virtue of conveying this counterfactual information that we should think of them as explanatory. For example, in the desert traveler case, in which the hole that *B* punctures rather than the poison that *A* inserts causes *T*'s death, the following counterfactuals are true: given that the actual situation is one in which *T* does not ingest poison, if *B* had not punctured, *T* would not have died, and if *B* had punctured, *T* would have died. Moreover, given that in the actual situation, *T* becomes completely dehydrated, he would have died regardless of whether or not *A* inserted poison. It is these facts about counterfactual dependence that account for our judgment that it is *B*'s puncturing and not *A*'s poisoning that explains *T*'s death. This is so despite the fact that both the poisoning and the puncturing are nomologically sufficient for *T*'s death—yet another illustration of the general point that identifying a condition on which an outcome depends (and which thus explains the outcome) is a very different matter from identifying a nomologically sufficient condition for the outcome. Similarly, in the falling boulder case, whether ducking occurs is counterfactually dependent on whether the fall occurs; hence, the fall of the boulder explains why the hiker ducks. Similarly, given the actual situation in which there is a fall, survival counterfactually depends on and is explained by ducking. But given the actual situation in which ducking occurs, survival is not counterfactually dependent on and hence is not explained by the boulder's fall.

Now consider some of the singular causal explanations discussed in chapter 4:

(5.8.1) Jones's latent syphilis caused his paresis.

(5.8.2) The impact of the knee on the desk caused the tipping over of the inkwell

as well as that stock favorite of philosophers:

(5.8.3) The short circuit caused the fire.

(5.8.1)–(5.8.3) do not explicitly invoke laws, and if the argument of chapter 4 is correct, they also do not implicitly invoke laws in the strong sense of law required by nomothetic models of explanation. Moreover, though we assumed, in the case of the desert traveler and boulder examples, the existence of deterministic background generalizations, if the argument of chapter 4 is correct, (5.8.1) and (5.8.2) do not as they stand assume or require (or at least need not be interpreted as assuming) that the causes they cite operate deterministically. Instead (5.8.1) and (5.8.2) are simply noncommittal on this point. Thus, for example, only 25 percent of those with third-stage syphilis

develop paresis, and though there may be some further conditions  $K$  such that when third-stage syphilis and  $K$  are present, paresis always follows, (5.8.1) does not claim that this is the case: the explanatory import of (5.8.1) does not depend on there being such an additional condition  $K$ . Similarly for (5.8.2) and (5.8.3).

In view of these facts, what can we say about the counterfactual import of (5.8.1)–(5.8.3), or about what they tell us by way of answers to what-if-things-had-been-different questions? Let us begin with the simplest sort of situation in which (5.8.1)–(5.8.3) might be used, a situation in which there is no causal overdetermination: no other cause of paresis besides syphilis, no other cause of ink spilling besides the knee impact, and no other cause of the fire besides the short circuit, are operative or waiting in the wings. Along with most other commentators who favor counterfactual accounts of causation, I claim that in this sort of case, a singular causal claim (or explanation) of form  $c$  caused  $e$  implies the following counterfactual: if  $c$  had not occurred, then  $e$  would not have occurred. In particular, this counterfactual will hold even if  $c$  is an indeterministic cause of  $e$ . For convenience, I call this a “not-not” counterfactual. Thus, (5.8.3) (or strictly speaking, (5.8.3) in conjunction with the information that other causes of the fire are absent) implies the following not-not counterfactual:

(5.8.4) If the short circuit had not occurred, the fire would not have occurred.

Similarly, (5.8.1) implies (again assuming the absence of other causes of paresis):

(5.8.5) If Jones had not contracted latent syphilis, he would not have paresis.

We can think of this counterfactual information just as it stands as conveying information about the answer to what-if things-had-been-different questions and as (at least minimally) explanatory for just this reason. (For what “minimally” means, see below). Thus, (5.8.1) may be regarded as explaining Jones’s paresis because it identifies a condition (latent syphilis) such that if this had been different (in particular, if it had been absent), this explanandum phenomenon would have been different (in fact, would not have occurred). Similarly for (5.8.2) and (5.8.3). As with other causal and explanatory claims, these counterfactuals should be understood as claims about what would happen if interventions were to occur; thus, (5.8.4) should be interpreted as claiming that if the short circuit had been caused not to occur as a result of an intervention, then the fire would not have occurred, and so on.

Before proceeding, let me comment on an interpretive issue that has been raised by several commentators (e.g., Bennett 1987). How exactly should we understand a phrase like “if the short circuit had not occurred” in (5.8.4)? It seems that there are, so to speak, a variety of different possible ways in which “the” short circuit might have failed to occur and that (5.8.4) may be true under some of these possibilities but not under others. Suppose that the actual

short circuit  $s$  occurred at a specific time  $m$  and reached a certain temperature  $T$ . Consider a short circuit  $s^*$  that occurs at a somewhat different time  $m^*$  and reaches a different temperature  $T^*$ . Will  $s^*$  be the same short circuit as the actual short circuit  $s$ ? If, as seems arguable for some values of  $m$  and  $T$ , the answer to this question is no, then one of the ways the (actual) short circuit  $s$  might fail to occur is if the different short circuit  $s^*$  occurs instead. But, if the antecedent of (5.8.4) is understood to include this sort of possibility—that is, if the nonoccurrence of  $s$  in (5.8.4) is understood to encompass the occurrence of  $s^*$ —then it is far from obvious that the counterfactual (5.8.4) is true. Suppose, for example, that the short circuit  $s^*$  reaches a much higher temperature than  $s$ . (Depending on the details of the case, there is an obvious sense in which a world in which  $s^*$  occurs might be closer to the actual world than a world in which no short circuit occurs at all.) Given this supposition, it is by no means obvious that the fire would have failed to occur, as (5.8.4) requires—or at least we are now confronted with the parallel question of what it means to talk of the nonoccurrence of “the” fire. Obviously, we need to find a way out of this difficulty if we are to associate counterfactuals like (5.8.4) with claims like (5.8.3).

There are complicated considerations having to do with event identity and event essences that some writers have brought to bear on this issue, but by far the most natural and straightforward solution, in my view, appeals to the notion of contrastive focus, introduced in chapters 2 and 3. According to this analysis, a singular causal claim like (5.8.3) has an implicit contrastive or “rather than” structure which spells out what is meant by the nonoccurrence of the cause and effect in (5.8.4). On the most natural way of understanding (5.8.3), it should be interpreted as claiming something like this:

(5.8.5) The contrast between the occurrence of the short circuit and an alternative situation in which no short circuit at all occurs causes (or causally explains) the contrast between the actual situation in which the fire occurs and an alternative situation in which no fire occurs.

Put slightly differently, (5.8.3) says that it was the occurrence of the short circuit rather than the occurrence of a situation in which no short circuit occurs that explains why the fire rather than no fire at all occurs. (5.8.3) does not claim, and it is presumably false, that:

(5.8.6) The contrast between the situation in which the short circuit occurs just as it actually did and an alternative situation in which a short circuit occurs at a somewhat higher temperature explains the contrast between the occurrence and nonoccurrence of the fire.

On this analysis, one of the functions of contrastive focus in a singular-causal explanation is to tell us what sort of situation we are supposed to envision when we consider a counterfactual like (5.8.4): the contrastive structure specifies what is meant by the nonoccurrence of the cause and the nonoccurrence of the effect. For example, if we interpret (5.8.3) along the lines of (5.8.5), then in envisioning a situation in which the short circuit does not

occur, we are supposed to think of a situation in which no short circuit at all occurs, not one in which a short circuit occurs that reaches a different temperature from the temperature reached by the actual short circuit. Similarly, we are to think in terms of a situation in which no fire at all occurs rather than in terms of a situation in which a fire occurs that is slightly different from the actual fire. In attributing the contrastive structure (5.8.5) to (5.8.3), we are in effect specifying or assuming that the counterfactual (5.8.4), understood along the lines described above, is the correct not-not counterfactual to associate with (5.8.3).

Of course, if the conditions governing the occurrence of the fire were sufficiently different, the alternative claim (5.8.6) might be correct. Suppose that we believe, for some reason, that the short circuit would not have caused a fire if the short circuit had reached some sufficiently higher temperature  $T^*$ . (Perhaps the material surrounding the short circuit will ignite only within a certain range of temperatures below  $T^*$ , or an automatic sprinkler system would have been triggered if the temperature had reached  $T^*$  but not if it had reached  $T$ .) In this case, it will be true that if  $s$  had not occurred, and  $s^*$  had occurred instead, the fire would not have occurred. But by the same token, it would now be natural to express the associated singular-causal claim as follows: the contrast between the occurrence of  $s$  and the occurrence of  $s^*$  explains the occurrence rather than the nonoccurrence of the fire. We now have a different singular-causal explanatory claim associated with a different counterfactual and a different set of contrasts. Which set of contrasts is intended when a claim like (5.8.3), which is not explicitly contrastive, is used, will of course depend on the details of the case.

The fact that in situations in which other causes are absent, singular-causal claims uncontroversially imply not-not counterfactuals is sufficient to show why such claims are explanatory, given the account of explanation sketched above: such claims tell us about a condition (the nonoccurrence of the cause) such that if it had occurred, the outcome being explained would not have occurred. Nonetheless, as we shall see below, singular-causal claims that are associated only with a single not-not counterfactual and not with any other counterfactuals are often explanatorily shallow or only minimally explanatory. For this reason, among others, it is natural to ask whether there are other counterfactuals besides not-not counterfactuals that may be plausibly associated with some or all singular-causal claims. As is well-known, David Lewis ([1973] 1986b) has a candidate for such a counterfactual: he contends that, given a true causal claim of form  $c$  caused  $e$ , it follows just from the fact that  $c$  and  $e$  occur, that the counterfactual

(5.8.7) If  $c$  were to occur,  $e$  would occur

is true, even if  $c$  is an indeterministic cause of  $e$ .

As explained in chapter 2, my view is that the counterfactuals that are relevant to causation (including those relevant to token-causal claims) should be understood as asserting the existence of reproducible relationships having to do with what would happen under repeatable interventions. Thus, (5.8.7)

would be true in the special case in which there is a reproducible deterministic relationship between  $c$  and  $e$  such that if a relevantly  $c$ -like event were to be introduced in similar circumstances on other occasions, an  $e$ -like event would occur. In cases of this sort, it will be appropriate to associate counterfactuals like (5.8.7) as well as not-not counterfactuals with singular-causal claims. Suppose, however, that the relationship between  $c$  and  $e$  is indeterministic. Then, in contrast to Lewis, I think that it is far from obvious that (5.8.7) is true if  $c$  and  $e$  happen to occur. For example, if a genuinely indeterministic coin toss  $t$  occurs and the event, call it  $h$ , of the coin coming up heads on that toss occurs, there seems to be little consensus that the counterfactual “If  $t$  were to occur, then  $h$  would occur” is true; instead, many people judge it to be false or else have no clear opinion about its truth value. For this reason, I will assume in what follows that in cases in which it is not justifiable to assume determinism, any additional counterfactual claims besides the not-not counterfactual “If  $c$  had not occurred,  $e$  would not have occurred” associated with singular-causal claims will take some other form besides (5.8.7).

In fact, there are at least two other sorts of counterfactuals that may be associated with many singular-causal claims, including those that are indeterministic, and that contribute to their explanatory import. First, there often will be an associated type generalization that is known to users of a singular-causal claim. For example, in the case of (5.8.1) (“Latent syphilis caused Jones’s paresis”), ordinary users will be aware of the following type-causal generalization:

(5.8.8) Latent syphilis causes paresis

and, as we will see below, will often appeal to this generalization in establishing that (5.8.1) is true. On my view, the generalization (5.8.8) may be plausibly seen as part of the background to (5.8.1) and as contributing to its explanatory import.

How should we understand (5.8.8)? At a minimum, (5.8.8) implies that there is a reproducible relationship between latent syphilis and paresis in the following sense:

(5.8.9) In the right circumstances, intervening to cause someone to have latent syphilis (i.e., manipulating the situation from one in which no latent syphilis is present to one in which latent syphilis is present) will *sometimes* change whether paresis occurs; in particular, it will change the situation from one in which the probability of paresis is 0 (because there is no syphilis) to one in which that probability is greater than 0. (Recall that at present we are confining ourselves to cases in which no other cause of  $e$  is present besides  $c$ .)

We might describe (5.8.8/5.8.9) as a “possible-cause” generalization: it tells us latent syphilis is among the possible causes of paresis. Similar possible-cause generalizations (impacts cause liquids to spill, short circuits cause fires) play a similar role in connection with (5.8.2) (the impact of the knee caused the spill) and (5.8.3) (the short circuit caused the fire). Moreover, as we shall see in

chapter 6, it is reasonable to take many possible-cause generalizations as saying something stronger than what has just been described (i.e., that the presence of a cause of type *C* (to which *c* belongs) will raise the probability of an effect of type *E* above 0 in *some* circumstances). Many possible-cause generalizations should be understood as claims that this manipulability relationship (5.8.9) will hold in *many different* circumstances or for *many different* interventions—as claims that the relationship is in this sense robust or invariant across such changes. For example, latent syphilis can cause paresis among people with a wide variety of backgrounds and in a wide variety of circumstances; its capacity to cause paresis is in this sense relatively invariant. As we shall see below, it is when the possible-cause generalization associated with a singular-causal claim is relatively invariant that we tend to find the causal claim satisfying from the point of view of explanation. Moreover, it is typically when a possible-cause generalization holds across some range of circumstances that we think of it as describing a means-ends relationship that is useful for purposes of manipulation and control.<sup>3</sup>

Given the singular-causal claim “*c* caused *e*,” an associated possible-cause generalization will be relatively invariant or robust in the sense just described if a family of counterfactuals of the following sort holds:

(5.8.10) Intervening to introduce a *c*-like event into a situation in which a *c*-like event would not otherwise occur changes the situation from one in which an *e*-like event would not have occurred to one in which an *e*-like event sometimes occurs in each of a relatively broad range of background circumstances.

I emphasize that I am *not* claiming that a generalization like (5.8.10) is associated with *every* true singular-causal claim. For some true singular-causal claims of the form “*c* caused *e*,” any associated type generalization that is framed in broadly the same vocabulary as the singular-causal claim (or at least any such generalization that users are aware of) will be relatively fragile and non-invariant. My claim is only that for many singular-causal claims, an associated generalization of the form (5.8.10) will hold, will be known to users, and will make a contribution to the explanatory import of the singular-causal claim. For example, in the case of (5.8.3) (“the short circuit caused the fire”), there will not only be an associated not-not counterfactual (5.8.4). It is also plausible to think of users of (5.8.3) as committed to counterfactuals of the form:

(5.8.11) If a short circuit were to occur in any one of a range of different background circumstances, then in those circumstance a fire will sometimes occur.

It might seem that such counterfactuals are so weak and uninformative that they can play no interesting role in causal investigation and understanding. Reflection on (5.8.3) shows that this assessment is mistaken. In addition to knowing (5.8.11), typical users of (5.8.3) will also know or believe another general piece of causal information that may seem to border on triviality, but in fact can be quite useful when combined with general information about the



other possible causes of fires. This is that (at least in ordinary experience) fires do not occur purely spontaneously: fires always have some cause. Moreover, although there are many possible causes of fires, not just any arbitrary occurrence can play this role, at least in the usual circumstances in which users of (5.8.3) find themselves. This allows one to engage in a characteristic pattern of eliminative argument that plays an important role in establishing singular-causal claims and the counterfactuals associated with them: one can reason that if the short circuit and fire occurred and no other causes of the fire were present or waiting in the wings, and the possible-cause generalization “Short circuits cause fires” is true, then the only remaining possibility for the cause of the fire is given by (5.8.3) and that the counterfactual (5.8.4) must be correct. In this way, one can make use of information like that in (5.8.11), which, though general, falls well short of knowledge of a deterministic law to establish (5.8.3).<sup>4</sup>

Given this connection between singular-causal claims of the form *c* caused *e* and the not-not counterfactual “If *c* had not occurred, then *e* would not have occurred” and counterfactuals of the form (5.8.10), we may think of them as answering a set of what-if-things-had-been-different questions and identifying a pattern of dependence in the following sense (which goes beyond just what is conveyed by the not-not counterfactual). Again confining ourselves to cases in which no other cause of *e* besides *c* is present, a singular-causal explanation identifies two possible outcomes, the occurrence or nonoccurrence of *e*, and two possible conditions in its explanans, the occurrence or nonoccurrence of *c*, and asserts that there is a pattern of dependence of these outcomes on these conditions. A change from a situation in which *c* occurs to one in which it does not would in those particular circumstances have produced a change from the actual situation in which *e* occurred to an alternative situation in which *e* would not have occurred. Moreover, even though the occurrence of *c* or a *c*-like event need not always be followed by the occurrence of *e* (or an *e*-like event), such an occurrence alters the situation from one in which *e* would not have occurred to one in which it will sometimes occur, and this will be true for a range of circumstances in addition to the actual circumstances.

On a *DN* (or, for that matter, any nomothetic) conception of explanation, it is hard to see why these counterfactual implications of singular-causal claims—implications that most commentators acknowledge to be an important part of their content—contribute to their explanatory import. As noted in chapter 4, philosophers committed to the *DN* model (and its modern descendants, such as unificationist approaches) have instead seen such explanations as conveying understanding in virtue of invoking a hidden structure involving nomologically sufficient (or, at least, probabilifying) conditions or, in the case of nomothetic models, a subsuming law. This, in turn, leads to the puzzle of how this information can be working so as to provide understanding even though it is epistemically inaccessible to ordinary users of such explanations. On the contrasting conception defended in this chapter, singular-causal explanations wear the source of their explanatory efficacy on their face: they explain not because they tacitly invoke a “hidden” law or statement of

sufficient (or probabilifying) conditions, but because they identify conditions such that changes in these conditions would make a difference for whether the explanandum phenomenon or some specified alternatives would ensue. The information about such conditions and the counterfactuals associated with them are epistemically accessible and nonhidden. This allows us to avoid the epistemological costs of hidden structure analyses of such explanations, while preserving the idea that there are important continuities between singular-causal and theoretical explanation.

I said above that singular-causal claims that are associated only with not-not counterfactuals, but not with other sorts of counterfactuals of the form (5.8.10), tend to be explanatory but only minimally so. In particular, I suggest that as a general rule we are more willing to regard a singular-causal claim as a satisfying explanation if the putative cause and effect are described in such a way that they fall under a relatively invariant possible-cause generalization. As an illustration, assume again that (5.8.3) the short circuit caused the fire and also assume, for the sake of argument, that the short circuit is (identical with) the most noteworthy event of the year. Compare (5.8.3) with

(5.8.12) The most noteworthy event of the year caused the fire.

Even if we regard (5.8.12) as true, it seems uncontroversial that, in comparison with (5.8.3), it is defective or unsatisfactory as an explanation: as Kim (1999, p. 17) puts it, “We have lost crucial explanatory *information*” in the move from (5.8.3) to (5.8.12). The analysis I propose traces this at least in part to the following consideration: although there is a not-not counterfactual associated with (5.8.12) that is true (at least under one natural interpretation), namely

(5.8.13) If the most noteworthy event of the year had not occurred, the fire would not have occurred

the associated possible-cause generalization and the counterfactuals corresponding to (5.8.10/11) are not true, and the associated possible-cause generalization is relatively noninvariant. That is, although “short circuits cause fires” is a true and indeed relatively invariant possible-cause generalization, there is no comparably invariant generalization that looks anything like

(5.8.14) Noteworthy events can cause (are possible causes of) fires.

(5.8.14) does not describe a reproducible means-ends relationship that one might exploit to produce a fire. “Introduce a noteworthy event” is not helpful advice for someone who wants to start a fire, and noteworthy events do not cause fires across a range of background circumstances in the way that short circuits do. Examples of this sort suggest that the entire explanatory content of a singular-causal claim such as “*c* caused *e*” is not carried just by the corresponding “not-not” counterfactual; whether the corresponding possible-cause generalization and the “might” or “can” counterfactuals it supports are true is also of explanatory relevance.

Finally, it is worth drawing attention to another set of counterfactuals that are relevant to the explanatory import of many singular-causal claims. These are counterfactuals that express what Lewis (2000) calls “influence,” which he characterizes as follows:

*c* influences *e* iff there is a substantial range  $c_1, c_2, \dots$  of different not-too-distant alterations of *c* (including the actual alteration of *c*) and there is a range  $e_1, e_2, \dots$  of alterations of *e*, at least some of which differ, such that if  $c_1$  had occurred,  $e_1$  would have occurred, and if  $c_2$  had occurred,  $e_2$  would have occurred, and so on. (p. 190)

Consider a paradigmatic case of causation: the impact of a rock, thrown by Billy, causes a bottle to shatter. Suppose that no other causes of the shattering are present. Then, not only will the shattering be counterfactually dependent on Billy’s throw, in the sense that if the throw had not occurred, the shattering would not have occurred, but a number variations or “alterations” in the time and manner of the shattering will be counterfactually dependent on alterations in the time and manner of Billy’s throw. For example, if Billy had thrown a bit earlier (or later), the bottle would have shattered earlier (or later); if the trajectory of the throw or the momentum imparted to the rock had been different in the right way but the rock had still hit and shattered the bottle, the details and manner of the shattering would have been somewhat different; and so on. In addition, we think that if certain aspects of this causal transaction had been different, the effect would not have occurred at all; for example, if Billy had thrown but the rock had not struck the bottle, then the bottle would not have shattered.

In the “new” theory of causation defended in Lewis (2000), he proposes to define causation as the ancestral of *influence*; the intuitive idea is that for *c* to count as a cause of *e*, it needn’t be the case that the *occurrence* of *e* depends counterfactually on the *occurrence* of *c*. It is sufficient that enough alterations of *e*, having to do with the time and manner of its occurrence, counterfactually depend on alterations of *c*, having to do with the time and manner of its occurrence. I am skeptical that this proposal will work as a general definition of causation,<sup>5</sup> but I agree with Lewis that in many (but not all) cases in which we make causal claims, we regard ourselves as committed to such influence describing counterfactuals. For example, in the absence of other causes of the shattering or esoteric Rube Goldberg devices waiting in the wings, anyone who believes that the throw caused the shattering will also believe that some range of changes in the time of the throw would have affected the time of the shattering. To the extent that such influence counterfactuals are true and known to and accepted by users of singular-causal claims, we may think of them as communicating information about more fine-grained and detailed patterns of counterfactual dependence than are communicated just by not-not counterfactuals. This information describes more fine-grained possibilities for manipulation and control; it is also explanatory information according to the general account of explanation advocated in this chapter, because it is information about how the explanandum phenomenon would have been different had the

explanans event been different in various ways (i.e., it is what-if things-had-been-different information). In other words, we should think of the explanatory information conveyed by singular-causal explanations as including not just information describing how the occurrence of one event depends on another, *but also as including information about the patterns of counterfactual dependence that establish that one event “influences” another.*

So far, I have been considering the counterfactual implications of singular-causal claims in contexts in which overdetermining causes or potential backup causes are absent. What about cases in which such causes are present? I will not attempt to provide a systematic treatment of such cases, but confine myself to recommending the general strategy followed in chapter 2 (which, in particular cases, might also be supplemented by some of the other counterfactual-based strategies in the literature, such as strategies that appeal to influence-describing counterfactuals of the sort considered immediately above). That is, the basic line I favor would see such claims as conveying information about patterns of counterfactual dependence (understood in terms of interventions and as involving “serious possibilities” in the sense described in 2.8), where these counterfactuals will sometimes have complex antecedents involving more than one intervention. The causal claims will be explanatory in virtue of conveying such information.

Consider a version of the desert traveler case, in which both the hole in the canteen and the poison are indeterministic rather than deterministic causes of death, with the hole as actual cause of the death and the poison as preempted cause. I have suggested that in the actual situation in which *T* does not ingest poison, it is false that if *B* had punctured, *T* would have died. Regardless of whether I am right about this, it remains true, as in the deterministic case, that, holding fixed that the actual situation is one in which *T* does not ingest poison, if *B* had not punctured, *T* would not have died. Moreover, given *T*'s dehydration, it is not true that if *A* had not inserted poison into the canteen, *T* would not have died. If instead the insertion of the poison was the actual cause of death and the puncturing a preempted cause, then a different set of counterfactuals would hold. It is also arguable that if the puncturing is the actual cause of death, then, at least in any realistic noncontrived case, we should expect that a set of influence counterfactuals describing the dependence of the timing and manner of death on the puncturing will hold that are quite different from the influence counterfactuals that would hold if in fact the insertion of the poison is the cause of death.

As another illustration, consider the classic case of symmetric overdetermination in which shots from two riflemen enter the victim's heart simultaneously, killing him instantly, where each would have caused immediate death on its own. We have already noted (section 2.7) that in such a case, a different set of counterfactuals will be true than in cases involving preemption such as the desert traveler example. In addition, it is also plausible that different fine-grained influence counterfactuals concerning the timing of death, pattern of damage to the heart, and so on will be true depending on whether one or both riflemen fire and when and how they fire. (If rifleman 1 does not

fire and rifleman 2 does, then the timing of death will depend on the time at which rifleman 2 fires; if rifleman 1 fires much later than rifleman 2 fires, then the timing of death will not depend on when or even whether rifleman 1 fires; and so on.) Once again, we can think of the original causal claim that both shots symmetrically caused the death as a way of summarizing or representing a complex set of facts about a pattern of counterfactual dependence (as answering a complex set of *w-questions*) and as explanatory for this reason. More generally, the general approach to explanation that I advocate will be vindicated as long as there is some distinctive set of counterfactuals that are associated with each of the different causal claims that hold in specific cases of preemption, symmetric overdetermination, and so on and which can serve, so to speak, as carriers for the explanatory content of these claims.

## 5.9 Explanations without an Interventionist Interpretation

My focus so far has been on explanations that exhibit patterns of counterfactual dependence having to do with what would happen under interventions. It is interesting to note that there are derivations that are sometimes regarded as explanatory but that exhibit patterns of dependence that are not naturally interpretable in this way. For example, it has been argued that the stability of planetary orbits depends (mathematically) on the dimensionality of the space-time in which they are situated: such orbits are stable in a four-dimensional space-time but would be unstable in a five-dimensional space-time.<sup>6</sup> Does the dimensionality of space-time explain why the planetary orbits are stable? On the one hand, this suggestion fits well with the idea that explanations provide answers to what-if-things-had-been-different questions on one natural interpretation: we may think of the derivation as telling us what would happen if space-time were five-dimensional, and so on. On the other hand, it seems implausible to interpret such derivations as telling us what will happen under *interventions* on the dimensionality of space-time. A similar issue arises in the context of mathematical reasoning. Mark Steiner (1978) argues that some mathematical proofs are explanatory and others are not, and that genuinely explanatory proofs are those that show us how the truth of some theorem depends on the assumptions from which the theorem is proved:

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

The idea that in an explanatory proof we see how the theorem changes in response to variations in other assumptions is, of course, very close to the idea

that explanations should answer *w-questions*, but again, the notion of an intervention has no application in mathematical proofs.

One natural way of accommodating these examples is as follows: the common element in many forms of explanation, both causal and noncausal, is that they must answer what-if-things-had-been-different questions. When a theory tells us how *Y* would change under interventions on *X*, we have (or have material for constructing) a *causal* explanation. When a theory or derivation answers a what-if-things-had-been-different question but we cannot interpret this as an answer to a question about what would happen under an intervention, we may have a noncausal explanation of some sort. This accords with intuition: it seems clear that the dependence of stability on dimensionality or the dependence of a theorem on the assumptions from which it is derived is not any sort of causal dependence.

### 5.10 An Objection: Manipulation and Understanding

It might seem that there is an obvious objection to the account sketched above. The objection is that one may have information relevant to the manipulation and control of an outcome, or relevant to answering a what-if-things-had-been-different question regarding it, and yet still have little understanding of why it occurred. This objection may seem most salient in the case of singular-causal explanations, although, if correct, it will apply quite generally. As an illustration, consider that one can know all of the information associated above with the singular-causal claim (5.8.1) “Jones’s latent syphilis caused his paresis”—that if he had not contracted syphilis he would not have developed paresis, that syphilis is a possible cause of paresis, and so on—and yet understand nothing about why syphilis causes paresis. Doesn’t this show that, contrary to what I have supposed, (5.8.1) is unexplanatory (see Dretske 1977 for essentially this argument)? Similarly, one can know that (5.8.3) the short circuit caused the fire without knowing why or what it is about short circuits that causes (or explains) fires. Again, one can know that (5.10.1) turning the key in the ignition of my car is a way of manipulating whether the motor is on or off, while understanding nothing about how this action produces this effect or how the motor works.

Moreover, a parallel objection can be raised in connection with theories or bodies of knowledge that are more “scientific” in character. According to the manipulationist account, so-called phenomenological laws or generalizations can figure in explanations as long as these support the right sorts of counterfactuals about what would happen under interventions, even if they tell us nothing about underlying mechanisms, processes, or constituents. For example, on the theory sketched above, I can appeal to the ideal gas law  $PV = nRT$  to explain why, when the temperature of a sample of gas enclosed in a fixed volume is increased, the pressure will increase as well. Knowing this explanatory information allows one in principle to manipulate the pressure and to answer a range of *w-questions* about it in conformity to the general pattern

exhibited in (5.1.3)–(5.1.4). Yet obviously, this information, in itself, tells me nothing about why an increase in temperature produces an increase in pressure. It completely omits the more detailed causal story provided in statistical mechanics, in which we learn that the gas is made up of a very large number of constituent molecules that collide and that transfer energy and momentum to one another in accord with the laws of Newtonian mechanics. This dynamical story tells us, in molecular terms, what pressure and temperature *are*, and in doing so explains why the gas laws and many other phenomenological generalizations concerning the behavior of gases hold. According to many philosophers (e.g., Salmon 1984), it is only this statistical mechanical account that is genuinely explanatory; by contrast, the ideal gas law is merely descriptive.

Let me begin with the point about singular-causal explanations. I agree, of course, that merely knowing that (5.8.1) latent syphilis caused Jones's paresis doesn't tell one *why* syphilis causes (or what it is about syphilis that causes) paresis. However, it is a mistake to conclude from this observation that (5.8.1) is entirely unexplanatory. (5.8.1) does explain something: namely, why Jones developed paresis. The objection under consideration in effect assumes that it is a general requirement that an explanation that appeals to *S* to explain *M* must also explain why *S*s cause (or what it is about *S*s that cause) *Ms*. This assumption entails that it is impossible to ever explain anything: we cannot appeal to the positions and masses of the planets and the Newtonian gravitational force law to explain the trajectories of the planets because this explanation does not explain why those masses have this effect or why gravity obeys this law, and so on. The way out of this difficulty is to distinguish, as I have above, between appealing to *S* to explain *M* and explaining why *S*s cause (or are lawfully followed by) *Ms*. If explanation is possible at all, it must sometimes be possible to successfully complete the former activity without completing the latter.

The examples under consideration are meant to motivate the claim that explanation involves something different from the provision of information relevant to manipulation or to answering what-if-things-had-been-different questions. In fact, however, more detailed reflection seems to vindicate rather than undermine the account I have been advocating. To see this, consider first what would be required for an explanation of why syphilis causes paresis. Presumably, what one would like to know is just what features of the brain and nervous system are damaged by syphilitic infection, the mechanisms or pathways by which this damage comes about, and how this leads to paresis. I claim that this is just further, more detailed information relevant to manipulation and control, information that provides answers to a finer-grained, more detailed, and wider-ranging set of what-if-things-had-been-different questions. This sort of information might be used, for example, to block or interfere with processes by which nervous system damage occurs after infection by syphilis: it is, in effect, information that shows us how, if various internal processes and mechanisms were changed, the damage to the nervous system and hence the degree or extent of paresis would change. Thus, once

one recognizes what would be required for an explanation of why syphilis causes paresis, it looks as though the case under discussion does not serve as a counterexample to the manipulationist account, but rather supports it. What is required for deeper explanation or understanding in connection with syphilis and paresis is not something different in kind from the sort of information on which the manipulationist account focuses, but rather more of the same: better, more fine-grained and detailed information about dependency patterns of the same sort exhibited by (5.1.3)–(5.1.4) that appeal to generalizations that can be used to answer a wider range of *w-questions*.

Similar remarks apply to explanations involving phenomenological generalizations such as the ideal gas law. As I see it, the ideal gas law allows us to answer a certain range of what-if-things-had-been-different questions, and in doing so explains. Statistical mechanics does not explain in virtue of doing something different in kind from this, but instead simply provides information that allows us to answer what, in some respects, is a wider, more detailed range of *w-questions*; hence the sense that in some respects, it provides deeper explanations. (For more on this subject and the “in some respects” qualification, see section 5.12.) The objection under consideration is thus correct in thinking that explanations like (5.8.1) and (5.10.1) are relatively shallow, but wrong to think that it follows that the manipulationist account is mistaken. Instead, as I try to show in more detail in subsequent chapters, the manipulationist account can be used to spell out just what this shallowness consists in.

This having been said, I should also explicitly acknowledge a closely related respect in which the theory presented above runs contrary to a widely shared idea about explanation. This is that successful explanation is at bottom a matter of getting fundamental ontology right; on this view, an explanation can't be satisfactory if it doesn't tell us what the basic objects or constituents are that make up the systems whose behavior we are trying to explain. According to this view, the statistical mechanics of gases is a satisfactory explanatory theory because it tells us what sorts of entities gases are made up of or composed of; by contrast, the generalizations of phenomenological thermodynamics have little or no explanatory import because they are noncommittal or unrevealing regarding the basic constituents of gases or compatible with fundamentally mistaken assumptions regarding those constituents.

The manipulationist theory assigns a more limited significance to correctness at the level of fundamental ontology. According to that theory, the heart of explanation has to do with getting dependency or manipulability relations right; what matters is not so much what things *are* intrinsically, so to speak, but rather their *relational* features and behavior: in particular, how changing some quantity, property, or feature possessed by a system will change some other quantity, and whether such patterns of dependence are correctly described by the theory. Of course, one cannot talk of changing some quantity or feature without having some conception of what is being changed, and to this extent, explanatory knowledge cannot be completely divorced from ontology. It is also the case, as the example of statistical mechanics again illustrates, that detailed knowledge of the behavior of constituents often carries with it more



detailed and fine-grained information about dependency relationships. Nonetheless, there are many examples of theories that are manipulatively successful, even though, from the perspective of later theories, they are fundamentally mistaken about just what entities are being manipulated. On the account of explanation I favor we should regard such theories as explanatory. A theory might, for example, correctly capture the dependency relations between a certain set of measured quantities and, hence, qualify as explanatory, even if it says nothing about or makes mistaken claims about intervening processes or mechanisms.

As an illustration, consider the heat diffusion equation

$$(5.10.2) \quad dT/dt = D\nabla^2 T$$

where  $T$  is temperature,  $t$  time, and  $D$  a diffusion constant. This tells us how the rate of change of temperature with time at every point depends on the spatial distribution of temperature as reflected in the Laplacian of  $T$  and the value of the diffusion constant. (5.10.2) says nothing about what heat or temperature are at the molecular level and can be (and was) used by investigators who mistakenly thought of heat diffusion as the diffusion of a continuous fluid. Nonetheless, according to the manipulationist account, (5.10.2) may be used to explain the rate at which temperature changes in various materials. Similarly, both eighteenth-century particle theories and nineteenth-century wave theories correctly answered a range of *w-questions* about how the behavior of light would change under the manipulation of reflecting and refracting media, under the conditions associated with the creation of various sorts of diffraction phenomena, and so on, even though such theories are, from our present perspective, fundamentally mistaken about the nature of light.

In this sense, the account of explanation presented above is less demanding about the need for explanatory theories to have a defensible “realistic” interpretation (in the sense of postulating only “real” entities) and assigns a more prominent role to “instrumental” success than many competing theories of explanation. I have argued elsewhere (Woodward, 2003a) that this is a virtue of the manipulationist account; contrary to the inclinations of many philosophers, such an account fits explanatory practice in many areas of science much better than more ontologically oriented alternatives.

## 5.11 Explanations and Omissions

I noted in chapter 3 that on a manipulationist account of causation, at least some omissions or preventings will count as causes. A parallel conclusion holds for the manipulationist theory of explanation: given a variable  $V$ , one value  $v$  of which represents an omission, if interventions that change the value of  $V$  away from  $v$  will change the value of some second variable  $Y$ , then the fact that  $V=v$  can figure in an explanation of the value of  $Y$ . Thus, a doctor’s failure to treat a patient can figure in an explanation of the patient’s

death. It seems to me that any acceptable theory should recognize some cases involving omissions as bona fide explanations.

Recently, examples involving causation by omissions or prevention, but with a more complex structure, have been explored in the literature on causation. Of particular interest are cases involving "double prevention" or "causation by disconnection" (see Hall, forthcoming; Schaffer 2000a). These are cases in which an event *A* prevents the occurrence of some second event *B*, which, had it occurred, would have prevented the occurrence of a third event *C*. As a result, *C* occurs. The question is whether we should think of the occurrence of *A* as a cause of or part of the explanation of the occurrence of *C*. Hall (forthcoming) offers the following example. Suzy is piloting a bomber and Billy is piloting a fighter as her escort. Billy shoots down an enemy plane that, had he not shot it down, would have shot down Suzy and prevented the bombing. Given Billy's action, Suzy successfully carries out the bombing. Is Billy's firing a cause of (or does it figure in the explanation of) the bombing? Here it may seem (at least to some) that our judgment is pulled in two directions. On the one hand, the bombing is counterfactually dependent on Billy's firing (when this counterfactual is understood in an interventionist, non-backtracking sense), and this inclines us to answer yes to the above questions. On the other hand, Billy's firing may seem to lack the right sort of connection to the bombing to count as a cause of it; for example, there are no intervening events that connect the firing to the bombing via a spatiotemporally continuous process, and there is no transfer of energy and momentum from the former to the latter.

Hall's own diagnosis of this example is that we operate with two distinct concepts of causation, one of which ("dependence") involves counterfactual dependence but does not require a spatiotemporally continuous process connecting cause and effect, and the other of which ("production") at least usually requires such a process but not counterfactual dependence. Billy's firing is a cause of the bombing in the dependence sense but not in the production sense. My own view is that we should resist this particular proliferation of "concepts" of causation and that, as the manipulationist account suggests, Billy's firing is straightforwardly a cause of the bombing.<sup>7</sup>

If double prevention were confined to unusual examples of the sort just described, it may seem that it matters little whether we decide to count it as a case of causation. In fact, however, there many examples of more "scientific" explanations that have a similar structure. This is particularly true in biology, where it is common for one event or process *C* to cause another *E* by interfering with some third process *F* that inhibits or prevents *E*. It would be a real defect in a theory of causation/explanation if it did not allow us to recognize such examples as cases in which *C* causes or figures in the explanation of *E*.

As an illustration, consider the lac operon model for *E. coli* due to Jacob and Monod, which was widely regarded as a seminal discovery in molecular genetics. When lactose is present in its environment, *E. coli* produces enzymes that metabolize it, but when lactose is absent, these enzymes are not produced. What determines whether these enzymes are produced? According to the

model proposed by Jacob and Monod, there are three structural genes that code for the enzymes as well as an operator region that controls the access of RNA polymerase to the structural genes. In the absence of lactose, a regulatory gene is active which produces a repressor protein which binds to the operator for the structural genes, thus preventing transcription. In the presence of lactose, allolactose, an isomer formed from lactose, binds to the repressor, inactivating it and thereby preventing it from repressing the operator, so that transcription proceeds. Biologists describe this as a case of “negative control.” Unlike “positive control,” in which “an inducer interacts directly with the genome to switch transcription on” (Griffiths, Miller, Suzuki, Lewontin, and Gelbart 1996, p. 550), the inducer in this case, allolactose, initiates transcription by interfering with the operation of an agent that prevents transcription. In other words, this is a case of “double prevention.” A causal relationship is clearly present between the presence of allolactose and the production of the enzymes, and the former figures in the explanation of the latter, but there is no transfer of energy from, or spatiotemporally continuous process linking, the two. A counterfactual theory of the sort outlined above is the obvious candidate for capturing this relationship. Allolactose induces or causes production of the enzymes because production would occur when it is introduced via a properly designed experimental manipulation and would not occur in its absence, again assuming a properly designed experiment in which some other inducer is not present.<sup>8</sup>

## 5.12 The Role of Pragmatics

In my discussion so far, I have defended the idea that explanations must answer what-if-things-had-been-different questions. However, even putting aside the remarks in 5.10, it seems clear that there are purported explanations meeting this condition that (at the very least) strike us as unsatisfying. Some purported explanations involving omissions are cases in point. It is true that if the stranger *X* had administered the antibiotic to the patient in the example described in section 2.8, the outcome would have been different: the patient would not have died. Nonetheless, most people would hesitate to regard *X*'s failure to administer the antibiotic as part of the explanation of why the patient died. It is also true that if a large meteor had struck my office just as I was typing these words, I would not have typed them, but again, we are reluctant to accept the failure of the meteor to strike as part of the explanation for my writing what I did.

Examples of this sort, as well as others considered below, raise an obvious question: Is it explanatory to cite just any variable so long as there is *some* change in the value of that variable that (when brought about by an intervention) is associated with *some* change in the outcome we are trying to explain? Or are there additional considerations that lead us to regard only certain possible changes in an explanandum (and associated with these, only some possible changes in candidate explanans variables) as of explanatory

interest or as having explanatory import? When we require that a successful explanation show how changes in the explanandum variable would occur under changes in the values of explanans variables, are any changes as good as any other for this purpose, or is information about certain changes more important or relevant than information about others?

One response to the examples with which I began this section is to exclude them by insisting that absences, omissions, and preventions can never figure in causal explanations. Even if we had a principled general distinction among absences, omissions, and preventions, on the one hand, and whatever contrasts with them ("presences"?) on the other hand, this strategy would be misguided. As the examples in the previous section illustrate, omissions and preventions can sometimes figure in genuine explanations. What we need is a more nuanced strategy that recognizes that some appeals to absences are explanatory and yet that also enables us to see why others are unsatisfying. It is at this point that the complicated and ill-understood set of considerations that go under the rubric of the *pragmatics of explanation* enter our story.

One kind of consideration that is broadly pragmatic has already been discussed: the notion of a "serious possibility." If a change in a purported explanans is associated with some corresponding change in an explanandum, but the change in the explanans is not a serious possibility, then the information that the explanandum will change under this change in the explanans is typically not regarded as explanatory, or at least the purported explanation is not seen as satisfying or relevant or what anyone was looking for. The examples immediately above illustrate this point.

Considerations having to do with what is a serious possibility are often, although not always, intertwined with another set of considerations that affect our judgment of whether an explanation is relevant or satisfying. As noted, when we ask for an explanation of some outcome, we often (perhaps even typically) have in mind accounting for the contrast between that outcome and a specific alternative or set of alternatives to that outcome, rather than all possible alternatives. Often, we are not interested in, and our request for explanation should not be taken as a request to account for, the contrast between the explanandum phenomenon and other alternatives outside this set of interest. In other words, there is a specific what-if-things-had-been-different question or range of such questions that we want answered. Broadly pragmatic considerations play a role in specifying which of these are of interest. For example, in the case of the doctor who fails to come to the aid of the patient *P* who dies, we will almost certainly be interested in the contrast between the actual situation in which *P* dies and similar, non-far-fetched alternative situations in which *P* lives, rather than in contrasts involving outlandish scenarios in which *P* lives. The most obvious candidates for non-far-fetched situations in which *P* lives are *actual* situations in which patients like *P* survive: situations in which, say, other patients in hospitals who are the responsibility of the same or similar doctors and who develop a similar fever have survived. Obviously, the stranger *X* fails to intervene both in the situation in which patient *P* dies and in these alternative actual situations in which other patients

live; *X*'s failure to intervene does not distinguish the former from the latter. What does distinguish *P*'s death from these alternatives is that *P*'s doctor fails to administer the antibiotic, whereas the doctors of the other patients usually or always do—hence our sense that citing the doctor's failure is more germane or satisfying from the point of view of explanation than citing *X*'s omission. Similarly, a request for an explanation of why I wrote certain words will usually be taken as a request for factors that distinguish the actual situation from an alternative in which I write something different, or (less naturally) as a request for a factor that distinguishes the actual situation from an alternative in which I continue to live but do something besides write. A failure to be struck by a large meteor will characterize all of these situations and will not distinguish among them. On the other hand, my desire to express one set of ideas rather than another may well distinguish the actual situation from alternatives in which I write something different and will be regarded as providing a satisfying explanation for just this reason.

Consider another example.<sup>9</sup> When Al drops by Burt's office and invites him to go for coffee, Burt sometimes but not always accepts. As it happens, both Al and Burt are fairly straight-laced. Thus

(5.12.1) If Al were to appear in Burt's doorway wearing a shirt covered with obscene slogans, Burt would refuse to join him for coffee.

But because Al is also straight-laced, he would never wear such a shirt. If we want to explain why Burt joined Al for coffee on some particular occasion, it seems unsatisfying to cite

(5.12.2) Al's failure to wear a shirt covered with obscene slogans.

Nonetheless, it is true that there is a change in this condition (a change from a situation in which Al does not wear the shirt to one in which he does), which would change whether Burt joins Al: the what-if-things-had-been-different condition is satisfied.

My diagnosis of this example is that a request for an explanation for why Burt joined Al for coffee on some specific occasion will usually not be taken as a request for a factor or variable that distinguishes this occurrence from just any possible alternative situation in which Burt declines, but rather as a request for a factor that distinguishes the actual situation from certain very specific alternative situations in which Burt declines. In the situation as described, a very natural candidate for the alternative situations of interest are those other actual situations, call them *O'*, on which Al asks and Burt declines. That is, when asked for an explanation of why Burt agreed to coffee on occasion *O*, it is very natural to take the question of interest to be: Why did Burt agree to coffee on this occasion *O* when he does not on the other occasions *O'*? Obviously, citing a factor like Al's failure to wear a shirt covered with obscenities does not distinguish the present situation *O* from the alternatives *O'*, because that failure is present in all the actual situations. Instead, one looks for some factor (such as the fact that Burt has no pressing deadlines or has some gossip that he wants to relay to Al) that occurs on the present occasion but that is absent

when Burt declines. I take this to be what underlies our sense that (5.12.2) is unsatisfying or inappropriate as an explanation of why Burt has joined Al.

On this analysis, the problem with (5.12.2) as an explanation of why Burt joins Al is that it fails to explain the explanandum that we want explained: namely, why Burt joins Al on this occasion but not on the other occasions  $O'$ . The case is thus like one in which I ask for an explanation of the deflection of starlight by the sun and you, misunderstanding me, take me to be asking for an explanation of the photoelectric effect and provide that instead. The explanation you provide does not explain what I want explained and is unsatisfying or inapt (as far as I am concerned) for this reason, but this is not to say that it is not a genuine explanation of the photoelectric effect.

If we think of the example involving Burt and Al as like this last example, then what both show is merely that what we want to explain—the particular explanandum we want to account for—often depends on our interests or on contextual or background factors. It is our interests that lead us to look for an explanation of why Burt joined Al on this particular occasion though he did not on the many other occasions  $O'$  on which Al asked, and not to care about finding a factor that distinguishes the present occasion from alternative, far-fetched scenarios in which Al shows up with obscene slogans.

Suppose that we change the example by imagining that some of the occasions  $O'$  on which Al invites Burt happen to fall on Tuesday, and that on those occasions Al does wear an obscene shirt as well as yellow socks, and that on other occasions  $O$  that are non-Tuesdays, he wears neither of these, and that Burt only joins him on occasions  $O$ . In this case, our interests may well shift in such a way that we are now interested in accounting for the contrast between Burt's behavior on those occasions  $O$  on which he joins Al and the alternatives  $O'$ , and our concern is whether it is Al's shirt, his yellow socks, its being Tuesday, or something else that accounts for this difference. The fact that Al's shirt contains no obscenities is now a plausible candidate for an explanation of the contrast of interest. A similar shift in what we are interested in explaining can also be induced by varying the example in a slightly different way: imagine that Clara has offered Al a considerable sum of money to wear a shirt with obscenities and that Al, being greedy as well as prudish, is greatly tempted, but in the end, after much indecision, declines to wear the shirt, asks Burt for coffee, and Burt agrees. Now the alternative scenario in which Al wears the shirt and Burt refuses to accompany him no longer seems far-fetched or not a serious possibility (it "almost happened"), and to the extent that we are interested in accounting for the contrast between the actual outcome and this alternative, citing Al's failure to wear the shirt will seem satisfying and to the point. On this view, both Burt's desire to exchange gossip with Al and Al's failure to wear the shirt with obscene slogans figure in genuine explanations; it is just that they figure in explanations of different explananda (why Burt joins Al on occasion  $O$  but not on the alternative occasions  $O'$ ; why Burt joins Al on occasion  $O$  but not on occasion  $O'$ ). In the case as originally envisioned, we were interested in an explanation of the first of these explananda and were instead given an explanation of the second.

Many philosophers advocate a quite different analysis of examples of this sort. They would describe the example as one in which whether Al's failure to wear a shirt with obscene slogans explains a fixed explanandum (whether Burt joins Al) depends on the interests of those involved. They conclude from such examples that explanation is a pragmatic notion in the sense that whether some candidate explanans *S* explains (i.e., bears an explanatory relationship to) some explanandum *M* depends on the interests of those who give or receive the explanation. On this view, pragmatics enters into the characterization of explanatory relationships themselves; that is, the same explanans *S* may or may not explain *M* depending on people's interests. This contrasts with the analysis that I favor, according to which the case under discussion is one in which people's interests shift from explaining explanandum *M'* to explaining the different explanandum *M*, and this requires a corresponding change in the explanans employed. On my analysis, interest relativity enters into what we explain but not into the explanatory relationship itself. What we try to explain depends on our interests, but it does not follow that for a fixed explanandum *M* and fixed explanans *E*, whether *E* explains *M* is itself interest-dependent. Obviously, it is not puzzling and no threat to the "objectivity" of explanation that the explanans *E* may explain *M* but a different explanans *E'* may be required to account for *M'*.

I will not try to provide a catalogue of all the various ways in which interests, context, or other background factors influence the kinds of *w-questions* we want answered. In many ordinary life cases, different people will have different interests and there will be a corresponding variability in the contrasts or differences they want explained. Thus, in Collingwood's (1940) well-known example, in which an auto accident occurs along a stretch of curved highway, the highway engineer may be interested in accounting for the contrast between the situation in which the crash occurs and alternative situations in which similar cars moving at similar speeds and with similarly competent drivers safely traverse other curves, and may conclude that the crash was caused by (or the explanation of the crash was that) the curve being (was) too tightly banked. By contrast, the judge or police officer may be more interested in explaining why this driver crashed along the curve whereas other drivers traversing the same curve have not, and in this case, such factors as the speed of the car and whether the driver was intoxicated will be of primary interest. Each investigator focuses on a specific (and different) what-if-things-had-been-different question. For example, for the engineer, the question of interest is this: Supposing that the speed, make of car, and driver were as they were in the actual situation, under what changes, if any, in the way the curve was built would the accident not have occurred? Again, rather than saying that the inadequate banking is an explanation of the accident for the engineer (given her interests) but not for the police officer (given his interests), it seems more perspicuous to say that both are explanations but are directed at accounting for different contrasts.

Both this example and the previous ones illustrate that the choice of explanandum we are interested in accounting for is often influenced by the range of

actual variations that happen to occur (in the present example, actual variation in drivers and curves): we are often more interested in accounting for actual rather than merely hypothetical or counterfactual variation (especially when the latter is regarded as unlikely to occur or far-fetched). However, as the example also illustrates, actual variation can occur along a number of different dimensions. Moreover, depending on the details of the case, the contrast between the actual outcome and merely hypothetical alternatives also may be of particular interest. For example, even (or perhaps especially) if all existing curves have the same design, the engineer may be quite interested in whether there is a possible but not actual variation in design that might have prevented the accident, especially if this variation is technologically possible and not prohibitively expensive.

In other cases, there will be less variation in interests in the relevant community of inquirers, and which what-if-things-had-been-different questions that investigators think are important or necessary to answer will be far more constrained. This is often the case in scientific investigation, especially when theories are already available that answer a wide range of *w-questions* regarding some body of phenomena. For example, given the prior existence of theories that answer a wide range of *w-questions* about electromagnetic phenomena, including questions about how the direction and intensity of electromagnetic fields will vary as the characteristics (charge, geometry, acceleration, etc.) of its source vary, any new electromagnetic theory that is regarded as explanatorily satisfactory will probably need to answer most or all of these questions and more besides. Given the success of prior theorizing, a theory that tells us how the field produced by a long straight wire would change as the current in the wire changes, but has nothing to say about the field that would be produced by sources with any other characteristics, will not be taken seriously. Instead, we demand an account that, in the manner described in section 5.1, shows how variations in the values of some same set of variables can be used in conjunction with the same laws or generalizations to answer a range of *w-questions* about all of these phenomena. In this case, as well as in many others, background knowledge and previous theoretical accomplishments sharply constrain the *w-questions* that are taken to be important to answer.

In another important and interesting kind of case, the grain or level at which theorizing takes place fixes or at least constrains the *w-questions* that are of interest.<sup>10</sup> Suppose that a mole of gas at temperature  $T$  and pressure  $P$  is confined in a cylinder with a movable piston. The piston is then withdrawn and the gas is allowed to diffuse into the new volume  $V'$  while a heat source maintains its temperature at  $T$ . Consider now the following (entirely impractical) strategy for explaining why the pressure exerted by the gas has changed to  $P'$ : one notes carefully the position and momentum of each of the  $6 \times 10^{23}$  molecules in the chamber immediately prior to the withdrawal of the piston (the initial microstate) and then explains the evolving energy and momentum of each molecule in terms of its initial state, the successive collisions it undergoes with other molecules, and the laws governing those collisions. The new pressure  $P'$  exerted by the gas is explained by aggregating the energy and



momentum transferred by each molecule to the walls of the container. For future reference, I will call this the microscopic strategy for explaining the behavior of the gas.

It is of course impossible, for a variety of reasons, to actually carry out this procedure. But even if a microscopic account of the sort envisioned were produced, there are important respects in which it would fail to provide the explanation of the *macroscopic* behavior of the gas we are looking for. This is because there are a very large number of different possible trajectories of the individual molecules in addition to the trajectories actually taken that would produce the macroscopic outcome that we want to explain. Very roughly, given the laws governing molecular collisions, one can show that almost all (i.e., all except a set of measure 0) of the possible initial positions and momenta consistent with the initial macroscopic state of the gas (pressure  $P$ , temperature  $T$ , and volume  $V$ ) will lead to a series of molecular trajectories such that the gas will evolve to the macroscopic outcome in which the gas diffuses to an equilibrium state of uniform density through the chamber at new pressure  $P'$ . Similarly, there is a large range of different microstates of the gas compatible with each of the other possible values for the temperature of the gas, and each of these states will lead to a different final pressure  $P'_i$ . It is an important limitation of the microscopic strategy that it does not, as it stands, capture or represent this information. That is, although the microscopic strategy would, if carried out, show that given the actual initial microstate of the gas, the gas "must" reach the final pressure  $P'$ , the strategy would not, as it stands, show which possible microstates would have led to the same final pressure  $P'$  and which would have led to a different final pressure. Because, on the view I advocate, explaining the final pressure of the gas is a matter of identifying conditions under which that pressure would have been different, the microscopic strategy omits information that is crucial to an explanation of the pressure. In this sense, an explanation that appeals to the initial macrostate of the gas—its initial temperature and volume—and a set of laws (like the ideal gas laws) relating such macrolevel variables and that shows how a range of changes in such variables would produce a different final pressure of the gas does a better job of showing what the final pressure depends on than the microscopic strategy. Although we could, of course, supplement the original microscopic account with an enumeration of which of the various microstates of the gas would lead to alternative values for the pressure and which would not, this is *additional* information connecting micro- and macrostate variables that is not plausibly regarded as provided by the original micro-level derivation of the pressure  $P$ .

The microscopic strategy does provide information about the details of the initial conditions and momenta of the molecules and their subsequent trajectories, which can be used to discriminate between an explanandum  $E$  specifying the detailed microstate of the gas at equilibrium and a range of alternative explananda  $E_i^*$  in which individual molecules have somewhat different positions and momenta. We can thus think of the microscopic strategy as explaining something: namely, why the actual explanandum  $E$  rather than the

alternatives  $E_i^*$  obtained. The problem, however, is that this is *not* the contrast (or set of what-if-things-had-been-different questions) in which we are interested when we ask why the gas has the macroscopic properties it does. Instead, what we want to know is why the gas has diffused uniformly throughout the chamber rather than being distributed in a nonuniform way, or why it exerts the final pressure  $P'$  rather than some other pressure  $P^*$ .

A similar point holds for many other explanations involving upper-level or macroscopic properties. Suppose that the price of oranges rises suddenly from  $P$  to  $P^*$ ; the explanation provided by economists is that there has been a freeze in Florida, the supply curve for oranges has shifted while the demand curve remains unchanged, and this has led to a new equilibrium with the higher price. Contrast this explanation with a (science fictional) derivation from fundamental physical laws and initial conditions  $S$  described in the language of fundamental physical theory of the physical state  $S^*$  that “underlies” (or realizes or is identical with) the highly distributed pattern of human behavior involved in oranges selling at price  $P^*$ . Again, there will be a large range of different physically characterized states besides  $S$  and  $S^*$  that are compatible with the initial price being  $P$ , the final price being  $P^*$ , and the supply and demand curves having the shape they do. There will be still other physical states such that if they had occurred instead, the final price would not have been  $P^*$ . Again, the contemplated derivation does not, at least in any very perspicuous way, convey this information: it does not identify those conditions under which the final price of the oranges would have been different and those under which it would have been the same. If we want to know what the final price depends on—that is, what changes in which variables would have led to a change in the final price—the explanation provided by economists does a better job of conveying such information than the physical explanation.

Both of these examples illustrate how the relationship between upper- and lower-level theories looks rather different when we think of explanation as a matter of tracing dependency relationships and answering *w-questions* rather than providing nomologically sufficient conditions. Different choices of variables for theorizing are associated with different ways of carving up nature into possible alternatives, answers to different *w-questions*, and hence different explanations.

### 5.13 More on Explanation and Change

A basic claim of this chapter has been that to explain an outcome, one must provide information about the conditions under which it would change or be different. As noted in 5.5, it follows that the generalizations that figure in explanations must be change-relating. This section explores an additional implication: if some putative explanandum cannot be changed or if some putative explanans for this explanandum does not invoke variables, changes in which would be associated with changes in the explanandum, then we cannot use that explanans to explain the explanandum, even if the relation

between the explanans and explanandum satisfies various other conditions standardly imposed in philosophical treatments of explanation. Both explainers and what is explained must be capable of change, and such changes must be connected in the right way.

Consider a derivation of the value of the gravitational constant  $G$  from the gravitational inverse square law, information about the masses of two objects, the gravitational force between them, and the distance separating them. This derivation satisfies the *DN* requirements, but is no explanation of why  $G$  has the value it does. Why not? On my view, it is because this derivation tells us nothing about the conditions under which  $G$  would change or take on a different value. From the point of view of Newtonian gravitational theory,  $G$  is a constant, which cannot be changed by changing other variables, either variables in the inverse square law or any other variables. Changing the distance between two objects or adding mass to one of them will change the gravitational force between them, but it will not change the value of  $G$ . To explain something, we must be able to think of it as (representable by) a *variable*, not as a constant or a fixed parameter. In contrast to competing accounts of explanation, the account sketched above explains, in a natural way, why this should be so.

Consider a more subtle example.  $N$  moles of gas are confined to a rigid container of fixed volume  $V$ . From the temperature  $T$  and pressure  $P$  in conjunction with the ideal gas law, we may deduce the value of  $V$ . Again, this derivation is no explanation of why the volume of the gas is  $V$ . My claim is that this is because the container is, by hypothesis, rigid: changing the values of  $P$  or  $T$  will not change the value of  $V$ . Now contrast this with a case in which the volume of the gas is allowed to vary: the gas is confined in a cylinder with rigid walls but with a piston that moves up and down in a vertical direction. A weight is placed on the piston, which exerts a net force  $F$  downward. The gas is heated to a new uniform temperature  $T^*$  and allowed to expand isothermally (i.e., its temperature is kept constant at  $T^*$ ) until the system is in equilibrium and the piston is at rest, at which point the gas occupies volume  $V^*$ . Here, in contrast to the previous case, one *can* explain why the gas occupies volume  $V^*$  by appealing to the ideal gas law, the new temperature  $T^*$ , and the new pressure  $P^*$  (itself fixed by the constraint that, at equilibrium,  $F = -P^*A$ , where  $A$  is the area of the piston). The difference between these two cases consists precisely in the fact that in the second case, but not in the first,  $V$  is variable or changeable. We can think of the second example as telling us something about the conditions under which the volume would have been different (if one changes the temperature  $T$  or the weight on the cylinder, one would change the value of  $V$ ), whereas the first example does not. According to the theory I have proposed, it is just this difference that accounts for the fact that we have an explanation of the value of  $V$  in the second case but not the first.<sup>11</sup>

Here is another illustration of the connection between explanation and change. In *The Nature of Selection* (1984) and in a more recent paper (1995), Elliott Sober argues that "natural selection explains the frequencies of traits in populations, but that selection cannot explain why individual organisms have

the traits they do" (1995, p. 384). Put slightly differently, natural selection provides what Sober calls *variational explanations* (of variation in the frequencies of traits in populations of organisms), but not *developmental explanations* (of the development of individual organisms). Sober illustrates this claim (and the contrast between these two kinds of explanations) by means of a nonbiological example. Suppose that to be a student in a certain classroom *R*, students must pass a test indicating that they read at the third-grade level. Some students *A* and *B* are admitted on the basis of this test, and others, *C* and *D*, are excluded. The use of this test amounts to a selection process, and the existence of this process explains, in one perfectly good respect, why it is true that

(5.13.1) All of the students in room *R* read at the third-grade level.

Nonetheless, Sober maintains, the existence of this selection process does not explain why this or that individual child (e.g., *A* or *B*) in the room *R* reads at the third-grade level, even though it follows from (5.13.1) that *A* reads at the third-grade level, *B* reads at the third-grade level, and so on. In just the same way, according to Sober, natural selection provides an explanation of why some trait *F* has a certain frequency within some population without explaining why individuals in the population have or develop *F*.

Sober's claim has been criticized (Neander 1988) and will strike some as paradoxical (how can one successfully explain why all the children in *R* read at the third-grade level while simultaneously failing to explain why any particular child in *R* reads at this level?), but the theory sketched above suggests that his claim is correct. On the manipulationist account, to explain an outcome we must cite conditions such that if they were to change in appropriate ways, the outcome being explained would also change. The selection process described above meets this criterion in connection with the explanandum (5.13.1) when this explanandum is properly understood (see below): if the selection process were different (if the requirement for entrance was instead that students read at the second-grade level or be over four feet in height), then (we may suppose) (5.13.1) would no longer hold. Different children would be in classroom *R* and some of them would not read at the third-grade level. However, changing the selection process in this way will not change the reading level of any particular child. If *C* reads at the second-grade level, changing the selection process so that those reading at this level are now admitted will have the result that he is now in *R*, but it will not alter his reading level. To do that, a different kind of intervention is required (e.g., a tutoring program that focuses specifically on *C*). This is why the selection process does not explain why any particular child reads at a certain level. Here again we see the plausibility of the idea that what will explain an outcome is closely bound up with what will change it. The air of paradox that surrounds the idea that an explanation of why (5.13.1) all the children in the room read at the third-grade level need not be an explanation of why some individual child reads at the third-grade level vanishes once one realizes that this is simply a reflection of the point that the conditions which, if changed, would change whether (5.13.1) is true, are not the conditions which,

if changed, would change the reading level of any particular child. We also see, once again, how it is physical dependency relations that matter for explanation and how deductive structure can fail to automatically mirror these. If I construct a derivation of (5.13.1) that appeals to some selection process  $S$ , I can use this same derivation, in conjunction with the additional information that some particular child  $C$  is in the room  $R$ , to construct a further derivation that has as its conclusion that

(5.13.2)  $C$  reads at the third-grade level.

Nonetheless, only the former derivation will be an explanation.

We should also note (as Sober himself points out) that all of this can be naturally expressed in terms of contrastive structure. In the case of the selectional or variational explanation described above, what one is interested in explaining is the contrast between the actual situation, in which the room is filled with certain children all of whom read at the third-grade level, and an alternative situation in which the room contains at least some different children, not all of whom read at the third-grade level. The reading level of any particular child is taken as given or fixed, and what is imagined to vary is the composition of the group in  $R$ , which, in virtue of the fact that it is taken to be capable of variation, is a potential object of explanation. By contrast, in the case of the developmental explanation, one is interested in the contrast between the actual situation in which this or that particular child (e.g.,  $A$ ) reads at the third-grade level and an alternative situation in which that very child reads at some different level. That is, the reading level of particular children is no longer taken as fixed, but as potentially varying and hence as an object of explanation. We thus see again the close connection between the what-if-things-had-been-different conception of explanation and the idea that explanations have contrastive structure.

## 5.14 Laws Again

In chapter 4, I remarked that a central and unresolved issue in the theory of explanation has to do with the role of laws: Do all explanations require appeal to laws? If so, what features must a generalization possess to count as a law? If, on the contrary, one can sometimes explain by appealing to generalizations that are not laws, what sort of features must such a generalization possess to count as explanatory?

The view that I have defended in this chapter is quite simple: what matters for whether a generalization is explanatory is whether it can be used to answer a range of what-if-things-had-been-different questions and to support the right sorts of counterfactuals about what will happen under interventions. (As we shall see in the next chapter, a generalization having these features must be *invariant* in the sense sketched in chapter 1: it must be stable under certain changes.) My argument in this and the following chapter is that it does not matter, independently of whether a generalization can be used to

answer *w-questions* (or whether it is invariant), whether we decide to classify it as a law or whether it possesses the other features traditionally assigned to laws by philosophers. That is, these other features matter only insofar as they have some connection with whether the generalization can be used to answer a range of *w-questions*. In fact, as already suggested, a generalization can lack many of the features traditionally assigned to laws by philosophers and can be dissimilar to paradigmatic examples of physical laws and yet still be such that it can be used to answer a range of what-if-things-had-been-different questions. For example, a generalization can hold only in a certain domain or regime and break down outside of this domain, and yet be such that we can use it to answer *w-questions* within the domain in which it holds. Similarly, a generalization can contain “positional” predicates or refer to particular objects or spatiotemporal locations and yet answer a range of *w-questions* and hence be used to explain.

This way of looking at matters allows us to avoid or bypass (or at least reconceptualize) a number of traditional disputes about the role of laws in explanation, and this in turn has a number of advantages. On the *DN* and other nomothetic models of explanation, to determine whether a putative explanation is genuine, we must first determine whether the generalizations to which it appeals include at least one law. However, as noted in chapter 4, the criteria for lawhood are highly controversial. Many proposed criteria are either unclear or fail to distinguish between laws and other sorts of generalizations or else apparently are not satisfied by many generalizations commonly regarded as laws. To the extent that this is so, and to the extent that it is true (as it surely must be) that lawhood has something to do with degree of similarity to paradigmatic examples of laws and this sort of similarity is itself vague and context-dependent, it is hard to avoid the conclusion that we are unlikely to find a sharp, generally accepted dividing line between laws and nonlaws.

Seen from this point of view, it is a real limitation in the *DN* model and nomothetic accounts of explanation generally that they require a sharp distinction between laws and nonlaws even though there is no generally agreed upon basis for this distinction. The account of explanation defended in this chapter avoids the need for such a sharp distinction by focusing instead directly on questions about whether a generalization can be used to answer a range of *w-questions*. In contrast to decisions about whether various generalizations count as laws, this is a matter that, as I try to show in chapter 6, we can often settle in a straightforward and unambiguous way.

In making this argument, I do not at all mean to suggest that the notion of a law of nature is empty or meaningless or that there are no laws in nature. On my view, there are indeed uncontroversial examples of laws, and these play an important role in some explanations. My claim is rather that, contrary to the views of many philosophers, we do not need to first settle the question of whether a generalization is a law before determining whether we can appeal to it to explain. From this point of view, the extensive attention in the philosophical literature devoted to the question of whether, say, commonsense

psychological generalizations or various generalizations of biology or economics are laws—a discussion usually predicated on the implicit or explicit assumption that it is only if such generalizations are laws themselves (or somehow tacitly invoke or point to laws) that we can appeal to them to explain—is misplaced. Attention ought to focus instead directly on the question of whether such generalizations can answer what-if-things-had-been-different questions and support counterfactuals about what will happen under interventions. I develop this theme in more detail in the following chapter.

## Invariance

### 6.1 Introduction

This chapter explores the notion of invariance. The guiding idea is that invariance is the key feature a relationship must possess if it is to count as causal or explanatory. Intuitively, an invariant relationship remains stable or unchanged as various other changes occur. Invariance, as I understand it, does not require exact or literal truth; I count a generalization as invariant or stable across certain changes if it holds up to some appropriate level of approximation across those changes.<sup>1</sup> By contrast, a generalization will “break down” or fail to be invariant across certain changes if it fails to hold, even approximately, under those changes. This characterization immediately raises two questions: Can we be more precise about what “stable” means? When we speak of stability under changes, what sort of changes matter? I explore these issues below. I want to begin, however, with some stage setting and intuitive motivation.

First, let me remind the reader of some issues, raised in earlier chapters, concerning the role of laws in explanation. We noted the absence of any consensus about the criteria that distinguish laws from nonlaws and the difficulties this posed for nomothetic accounts of explanation. We also noted that the so-called special sciences were full of apparent explanations that appealed to generalizations that fail to meet many of the standard criteria for lawhood. Given the assumption that laws are required for successful explanation, this generates a dilemma: either much of what we call “explanation” in the special sciences isn’t really explanatory or, alternatively, despite appearances to the contrary, explanatory generalizations in the special sciences do qualify as laws. Most philosophers have embraced the second horn of this dilemma, arguing that the generalizations of the special sciences are indeed laws, but laws of a special sort: *nonstrict*, *qualified*, or *ceteris paribus* laws. A great deal of energy is then expended trying to show how a generalization can be nonstrict or *ceteris paribus* and still qualify as a law.<sup>2</sup>

In my view, the appeal of this strategy lies not in its inherent plausibility, but rather in the difficulty of formulating a defensible alternative to nomothetic conceptions of explanation. The problem is that we lack a generally accepted positive theory of how generalizations function in explanations except as laws. In the absence of such a theory, even philosophers who are



well aware of the great distance between generalizations figuring in explanations in the special sciences and paradigmatic laws feel forced to assimilate the former to the latter. If we want to vindicate the idea that generalizations like Mendel's can be used to explain and the only possibilities are that such generalizations are either laws or completely accidental and nonexplanatory, we seem to have no alternative but to try to shoehorn them into the category of laws, however awkward the fit may be.

One of my goals in this chapter is to suggest a new way out of this dilemma. The standard way of thinking about laws suggests there are just two, mutually exclusive possibilities: either a generalization is a law (in the sense of satisfying at least many of the standard criteria for lawfulness) or else it is purely accidental. However, most explanatory generalizations in the special sciences do not fit comfortably into either of these two categories. We need a new way of thinking about generalizations and the role they play in explanation that allows us to recognize intermediate possibilities besides laws and accidents and to distinguish among these with respect to their degree or kind of contingency. This account should also allow us to understand how a generalization can play an explanatory role even though it holds only within a certain domain or over a limited spatiotemporal interval and has exceptions outside of these. I will try to show that the notion of invariance meets these requirements. In contrast to the standard notion of lawfulness, invariance is well-suited to capturing the distinctive characteristics of explanatory generalizations in the special sciences. As we shall see, this is in large measure because whether a generalization is invariant is surprisingly independent of whether it satisfies many of the standard criteria for lawfulness. A generalization can be invariant within a certain domain even though it has exceptions outside that domain or holds only with a limited spatiotemporal region. Moreover, unlike lawfulness, invariance comes in gradations or degrees.

None of this is to deny that there are generalizations that are legitimately regarded as laws of nature (the field equations of General Relativity, Maxwell's equations) and that play an important role in the explanation of many different phenomena. However, on the view I defend, such laws are simply generalizations that are invariant under a particularly wide range of changes and interventions. In general, it is the invariance of laws rather than whether, independently of this, they satisfy the standard criteria for lawfulness that endows them with explanatory import. Moreover, a generalization can be stable under a much narrower range of changes and interventions than paradigmatic laws and yet still count as invariant in a way that enables it to figure in explanations.<sup>3</sup>

Part of the appeal of the notion of invariance, then, is that it promises a more satisfactory treatment of the generalizations of the special sciences than the standard law versus accident framework. Moreover, ideas defended in earlier chapters provide additional motivation for assigning a central role to invariance in characterizing explanatory and causal relationships. Recall the account in chapter 5: this requires that the generalizations appealed to in each of (5.1.3)–(5.1.4) (the explanation of the field created by a straight wire, the

explanation of the price-increasing and output-restricting behavior of monopolies) are such that they can be combined with a range of possible changes in initial and boundary conditions to yield a range of different explananda or to answer a range of different *w-questions*. For these explanations to work in this way, it must be the case that these generalizations persist or continue to hold as we change the system whose behavior we are trying to explain. For example, the generalization (Coulomb's law) appealed to in (5.1.3) must be such that it holds in the case of the system whose behavior we are trying to explain (the electromagnetic field produced by a long carrying wire), but (as we have seen) it must also be such that it would continue to hold if the wire were twisted into a solenoid or into various other shapes when these changes are produced by interventions. If Coulomb's law did not have this feature, it could not be used to answer a range of *w-questions*. As noted above, Coulomb's law would also continue to hold across other sorts of variations in background conditions that are not explicitly mentioned in (5.1.3) because they are understood to be irrelevant—across changes in spatial and temporal location, for wires of different colors or made of different conducting materials, regardless of whether the wire was set up by experimenter *A* rather than experimenter *B*, and so on.

A similar conclusion holds regarding the other generalizations figuring in explanations discussed in chapter 5: they too must be appropriately invariant. Consider the generalization

$$(5.6.1) \quad V = b_0 + b_1 S + U$$

relating vote difference (*V*) to editorial slant (*S*) by the *Manchester Union Leader*. For this generalization to answer a range of *w-questions* showing how voting patterns change in response to editorial endorsements, it must (at least) remain stable under some interventions that change the value of *S*.

We thus arrive at the result that it is built into the theory developed above that explanations must appeal to generalizations (or laws or descriptions of dependency relations) that are invariant under some class of changes; in particular, the generalizations cited in an explanation must, at least, be invariant under some class of interventions on the initial or boundary conditions cited in the explanation.

A similar result follows if we think in terms of the manipulability theory of causation defended in chapter 2 and the idea that causal relationships are relationships that are exploitable in principle for purposes of manipulation and control. Suppose that we observe an association or correlation between *C* and *E*. If every intervention that changes the value of *C* disrupts any correlation between *C* and *E* (i.e., *C* and *E* become uncorrelated under this intervention), it will not be possible to use *C* to control or manipulate *E*, and we will not regard the relationship between *C* and *E* as causal. If, on the contrary, the association between *C* and *E* continues to hold (i.e., is invariant) under at least some interventions that change *C*, then (if these changes in *C* are associated with different values of *E*) we may be able to avail ourselves of the stability of this relationship to produce changes in *E* by producing changes in *C*.

Moreover, other things being equal, the greater the range of such changes under which the relationship is invariant, the more likely it is to be effectively exploitable for purposes of manipulation and control.

These considerations suggest that the notion of an intervention is crucial to characterizing the kind of stability that matters for whether a relationship is causal or can play an explanatory role: invariant relationships must at least be stable under interventions, although they may also and typically will be stable under other sorts of changes that are not interventions. The characterization of invariance below reflects this idea.

Yet another reason for taking invariance seriously is that one finds the idea that causal or lawful or explanatory relations must be invariant relationships in many areas of science. As we shall see in chapter 7, in the econometrics or causal modeling literature, the notion of a causal or explanatory relationship is closely connected with the notion(s) of structural or autonomous relationships, and these in turn are just relationships that are invariant in various ways. A typical (although, I suggest below, overly strong) characterization of this connection between causation (conceived along manipulationist lines) and invariance is Kevin Hoover's claim, quoted in chapter 2, that "causal relations are invariant to attempts to use them to control . . . effects" (1988, p. 66).

One finds a similar emphasis on invariance in connection with physical laws. An important constraint, emphasized by physicists but largely ignored in philosophical discussion, is that fundamental laws are expected to satisfy various sorts of symmetry requirements. When given a so-called active interpretation, these symmetry requirements are just invariance requirements: they amount to the idea that laws must describe relationships that remain stable or invariant under certain kinds of changes or transformations in the systems to which they apply. For example, one expects fundamental physical laws to be invariant under spatial and temporal translations, under Lorentz transformations, and so on. If causal explanation demands invariant relationships and if laws play an important role in (some) explanations, it is not surprising that laws are expected to satisfy these and other invariance requirements. We thus have the beginnings of a story, rooted in the idea that explanation requires invariant relationships, about why laws are expected to satisfy various symmetry requirements—something that is left unexplained on most philosophical accounts of lawfulness.<sup>4</sup>

In addition to the connection with symmetry requirements, the notion of invariance also provides a natural way of explicating one kind of "necessity" that laws possess and that merely accidental generalizations do not, a point emphasized in Skyrms's (1980) discussion of resiliency. This is the necessity that laws possess in virtue of describing relationships that, at least within a certain range, are immutable or unconditional in the sense that there is nothing we or nature can do that will make them fail to hold. At least part of what is implied by the claim that, say, the gravitational inverse square law is a genuine law is that this generalization would continue to hold under a very large range of possible changes, under changes in variables (mass, distance) that figure in this generalization and variables that do not. As Skyrms

notes, this sort of immutability or unconditionality is just another name for invariance.

There are also a number of other recent philosophical discussions of laws and causation that appeal to notions bearing a family resemblance to invariance. In addition to Skyrms's notion of resiliency, these include Michael Redhead's notion of robustness (1987), and Sandra Mitchell's notion of stability (1997, 2000). We may think of each of these notions as attempting, at least in some respects, to get at the same set of intuitions about how causal, lawful, and explanatory relationships must be stable relationships that I explore below.

However, although invariance is broadly similar to notions like robustness, my approach differs from other treatments of invariance-related notions in several respects. A common tendency in recent discussion is to think of invariance as requiring stability of a relationship under *all* possible changes. According to this conception, if a relationship fails to be stable under any possible change, then it is not invariant at all. By contrast, I urge that we should relativize the notion of invariance and recognize that a relation can be invariant under some changes and interventions but not under others. The picture I advocate involves both a threshold and a continuum. Stability under at least some interventions is necessary for a relationship to count as invariant. Generalizations that are not stable under any interventions at all are noninvariant: they fall below the threshold for invariance. In addition, however, we can distinguish among those generalizations that are invariant under at least some interventions with respect to the range or kind of interventions and other sorts of changes under which they are invariant. When we say that a relation is invariant, we have said something incomplete; we need to specify under which changes the relation is (or is not) invariant. This relativized notion of invariance is standard in econometrics and elsewhere in science and provides a better framework for thinking about causal and lawful relationships than an unrelativized, absolute notion. Among other things, the relativized notion allows us to express the idea that relationships differ in the "size" or "importance" of the range of interventions over which they are invariant. I argue that, other things being equal, relationships that are more invariant (and hence more useful for purposes of manipulation and control) provide better explanations. Fundamental physical laws that are invariant under a wide range of changes are at one end of this continuum; generalizations such as the regression equation (5.6.1) are closer to the other end.

The account I propose differs from many other treatments of invariance-related notions in the philosophical literature in another respect as well. We noted in chapter 2 that causal claims of the form

(6.1.1) *Cs cause Es*

are very nonspecific: they claim only that there is some intervention on *C* that in some circumstances will change *C* but say nothing about exactly how or when changes in *C* will change *E*. Such claims are noncommittal about which specific functional relationship, if any, links *C* to *E*. Discussions of invariance-related notions have often focused exclusively on claims of form (6.1.1) and

have asked whether there is some single universal “invariance condition” that is necessary and/or sufficient for any such claim to be true. By contrast, my view is that whenever a causal claim linking  $C$ s to  $E$ s is true, there must be *some* relationship connecting  $C$ s and  $E$ s that is stable or invariant under some range of interventions on  $C$ , but exactly which relationship is invariant and under which range of interventions will vary with the content of the causal claim in question. In my view, there are a number of different relationships on which we may legitimately focus when we talk about invariance and a number of different possibilities regarding the changes under which those relationships may be stable that are relevant to assessing invariance. Different sorts of causal claims will be associated with different sorts of claims about which relationships are invariant and about the changes under which they are invariant. We can spell out the content of a causal or explanatory claim by making this precise.

As an illustration, suppose that  $C$  and  $E$  are quantitative variables and the hypothesized relationship between them is a specific functional relationship, for example:

$$(6.1.2) \quad E = aC.$$

We may then ask whether this specific functional relationship is stable under interventions that change the value of  $C$ , as well as other sorts of changes. Obviously, however, stability of (6.1.2) cannot be a necessary condition for *all* claims of the form (6.1.1) to be true, because the relationship between  $C$  and  $E$  can be causal even though it is nonlinear or probabilistic rather than deterministic. Stability of (6.1.2) or of any other specific functional relationship under some range of interventions on  $C$  is not a universal necessary condition for causation, although it is an appropriate necessary condition for the very specific claim (6.1.2) to correctly represent a causal relationship.

Probabilistic theories of causation provide another illustration. Suppose that  $C$  and  $E$  are dichotomous random variables, taking the values 0 and 1, with a well-defined joint probability distribution. Obviously, one cannot represent causal relationships between such variables by means of linear relationships like (6.1.2), and thus one cannot capture what it is for a relationship between dichotomous variables to be causal by asking whether they are connected by an invariant relationship of the form (6.1.2). However, there are many other possible candidates for invariant relationships involving  $C$  and  $E$ . If  $C$  is the only factor that is causally relevant to  $E$ , then the existence of a causal relationship between  $C$  and  $E$  may show itself in the fact that the conditional probabilities  $P(E = 1/C = 1)$  and  $P(E = 1/C = 0)$  are not equal and are invariant under some range of interventions that change the value of  $P(C)$ . Under such conditions, one can manipulate  $P(E)$  by intervening on  $P(C)$ , and hence it makes sense, on the connection between manipulation and causation advocated above, to talk about a causal relationship between  $C$  and  $E$ . If  $E$  has multiple causes  $C_1 \dots C_n$ , the existence of a causal relationship between  $C_i$  and  $E$  may show itself in the invariance of the conditional probabilities  $P(E/C_1 \dots C_n)$

under interventions that change  $P(C_i)$ . Alternatively, it might be the case that these conditional probabilities are not invariant under interventions on  $P(C)$  but that the inequality  $P(E = 1/C = 1) > P(E = 1/C = 0)$  is invariant or continues to hold under some range of interventions on  $P(C)$ ; that is, changing the value  $C$  for some units in the population of interest from 0 to 1 always increases the probability of  $E = 1$  for those units, even though the numerical values for the conditional probabilities are not stable for different units or for different levels of  $P(C)$ . In this case too, one can manipulate the value of  $E$  (or at least the frequency with which different values of  $E$  occur) by manipulating the frequency of different values of  $C$ , and hence it will be appropriate to think of  $C$  as a cause of  $E$ , although of course, the relationship that is invariant will be different from the previous deterministic case. Chapter 7 describes in more detail a number of “invariance conditions” that are appropriate in probabilistic contexts.

Whatever the merits of the invariance conditions just described for probabilistic contexts involving dichotomous variables, we cannot apply them as stated to deterministic functional relationships involving real-valued variables like the gravitational inverse square law

$$(6.1.3) \quad F = Gm_1m_2/r^2.$$

Among other things, it is not clear that it is appropriate to think of the variables figuring in this relationship as random variables governed by a well-defined joint probability distribution. Instead, in this sort of case, as well as in others involving fundamental physical laws, the appropriate notion of invariance will be like that invoked in connection with (6.1.2), that is, stability of the specific functional relationship (6.1.3), over various sorts of changes. Again, different sorts of causal and nomological claims should be expected to satisfy different sorts of invariance conditions. For the most part, my focus in this chapter is on issues having to do with the invariance of specific functional relationships, with some brief comparative remarks about invariance in probabilistic contexts.

## 6.2 Invariance Characterized More Precisely

With this as background, let me turn to the task of characterizing the notion of invariance more precisely. I begin with a distinction between two sorts of generalizations; both conform to the “all As are Bs” framework often employed by philosophers, but differ in an important way. Some generalizations are change- or variation-relating, in the sense that they purport to describe a relationship between changes or variations in the value of one or more variables and changes or variations in the values of another variable. Focusing for concreteness on the two-variable case, a generalization  $Y = F(X)$  relating variables  $X$  and  $Y$  will be change-relating if and only if both of  $X$  and  $Y$  can take each of at least two values,  $x_1$  and  $x_2$  (where  $x_1 \neq x_2$ ), and  $y_1$  and  $y_2$  (where  $y_1 \neq y_2$ ), respectively, where  $y_1 = F(x_1) \neq y_2 = F(x_2)$ . It is important to realize

that both generalizations describing causal or lawful relationships and generalizations describing correlations that are "accidental" or "noncausal" can be change-relating. The generalization (6.1.3)  $F = Gm_1m_2/r^2$  is a change-relating law: it describes how variations in the magnitudes of either of two masses  $m_1$  or  $m_2$  or the distance  $d$  between them is associated with variations in the value of the gravitational force they exert on each other. The generalization describing the correlation between the value of a barometer reading  $B$  and the value taken by a variable  $S$  representing whether or not a storm occurs is also change-relating, even though that generalization does not describe a causal relationship between  $B$  and  $S$ . Indeed, any generalization describing a correlation (or pattern of association) will be change-relating, for anything that counts as a correlation must tell us how variations in the value of one variable are associated with variations in the value of another.

By contrast, other generalizations, including both some "accidental" generalizations and some laws, are not naturally interpreted as change-relating. Such generalizations say that some feature  $F$  is shared by all members of some class  $A$ , but do not say anything about whether  $F$  is or is not possessed by individuals that are not members of  $A$ . They don't link variations in the value of one variable to variations in the value of another. Examples include generalizations such as:

(6.2.1) All mammals have elastin in their arteries

and

(6.2.2) All igneous rocks have mass greater than zero

If we think of  $M$ ,  $E$ ,  $I$ , and  $Z$  as variables that take the values 0 or 1 depending on whether some object is or is not a mammal, has elastin, and so on, then these generalizations tell us what would happen when  $M$  and  $I$  take one of these values = 1, but do not purport to tell us what would happen under conditions in which  $M$  or  $I$  takes its other possible value = 0. (6.2.1) and (6.2.2) are not change-relating in the sense that they tell us how each of several different values of one set of variables is linked to each of several different values of other variables.

Also included in the general category of non-change-relating generalizations are generalizations that, to put it loosely, do not tell us how one set of changes is associated with another, but that certain changes are impossible. Consider the generalization

(6.2.3) No material object can be accelerated from a velocity less than that of light to a velocity greater than that of light

(6.2.3) is generally regarded as a law, but, as argued in chapters 2 and 4, it seems misguided to think of it as describing a relationship linking variations in the values of two variables. Insofar as the notion of a material object in (6.2.3) contrasts with anything at all, it presumably contrasts with the notion of a pseudo-process in the sense of Salmon (1984). Whether or not the argument of chapter 2 that there is no well-defined notion of changing a material object

into a pseudo-process or vice versa is correct, it seems clear that (6.2.3) says nothing at all about the possible velocities of pseudo-processes or other non-material objects. Rather than telling us, as the gravitational inverse square law does, how changes in one set of variables will produce changes in another set of variables, (6.2.3) tells us, in effect, that there are no physically possible changes in material objects that will produce a change from subluminal to superluminal velocities.

As another illustration, consider again the generalization

(6.2.4) All men who take birth control pills regularly fail to get pregnant.

There are at least two ways of understanding this generalization. First, on a literal reading, (6.2.4) does not even purport to be change-relating. Taken literally, (6.2.4) does not claim that changes in whether or not men take birth control pills are correlated with changes in whether or not they become pregnant, but only that male pill takers do not get pregnant. It says nothing about whether males who do not take birth control pills will get pregnant. Second, there is a nonliteral interpretation of (6.2.4) according to which it implicitly claims that the correlation just described does hold. Under this interpretation, (6.2.4) is false, because the claimed correlation fails to hold: whether or not a male takes birth control pills is not correlated with whether he becomes pregnant.

Obviously, if a generalization is to satisfy the what-if-things-had-been-different condition on explanation, it must be change-relating: it must be the case both that there is a well-defined notion of intervening on the independent variables in the generalization and that, under some such intervention, the value of the dependent variable will change. Because my interest is in the features that a generalization must possess if it is to figure in explanations, the notion of invariance on which I focus will apply only to change-relating generalizations. There is a legitimate notion of invariance that applies to generalizations that are not change-relating (as explained at the end of this section), but this notion is not at the center of my discussion.

Before continuing, it is worth commenting briefly on the implications of the discussion so far for the common philosophical practice of representing the structure of laws as universally quantified conditionals of the form

(6.2.5) All *As* are *Bs*.

A number of commentators have worried that such a representation leaves out important features of laws. We can now see at least one respect in which this worry is well-founded: the “All *As* are *Bs*” representation fails to represent the change-relating character of many laws. Both generalizations that are change-relating and those that are not will, if they hold universally, fit the representation (6.2.5), but if the argument of previous chapters is correct, only the former figure in causal explanations or represent causal relationships. A claim like (6.2.5) does not, as it stands, tell us even whether *A* and *B* are correlated, because it is compatible with *B* always occurring both in the presence of non-*As* and in the presence of *As*. By contrast, one of the virtues of



an equational representation is that it allows us to represent the change-relating character of generalizations.

Among change-relating generalizations, it is useful to distinguish several sorts of changes that are relevant to the assessment of invariance. First, there are changes in the *background conditions* to the generalization. These are changes that affect other variables besides those that figure in the generalization itself. For example, in the case of the gravitational inverse square law (6.1.3)  $F = Gm_1m_2/r^2$ , changes in the color of the masses or their electrical charge or the Dow-Jones Industrial average will count as changes in background conditions. Of course, the inverse square law would continue to hold under changes in such conditions.

Second, there are changes in those variables that figure explicitly in the generalization itself, for example, in the case of (6.1.3), mass and distance. Such changes divide into two subcategories. First, there are changes that result from an intervention on variables figuring in the generalization. The gravitational inverse square law is invariant under many such changes: it continues to hold under a wide range of interventions that change the distances between gravitating masses or the magnitudes of the masses themselves. Second, there are changes in variables figuring in a generalization that do not involve interventions; for brevity, I call these *non-I-changes*. An example would be a process that changes, at the same time, the mass of some object and its distance to a gravitating source. Of course, the inverse square law will continue to hold under many such non-I-changes.

The stability of a generalization under at least some interventions on variables that figure in it, as opposed to its stability under other sorts of changes (either changes in background conditions or non-I-changes), plays a privileged role in determining whether it describes a causal/explanatory relationship. Let me begin with stability under changes in background conditions. Any generalization, no matter how "accidental," nonlawful, noncausal, or unexplanatory, will continue to hold under *some* changes in background conditions. For example, the generalization describing the relationship between the barometer reading  $B$  and the onset of a storm  $S$  will be stable under the following changes in background conditions: changes in the position of Mars, the leadership of Russia, and the price of tea in China. Similarly, the paradigmatically accidental generalization

(6.2.6) All the coins in Bill Clinton's pocket on January 8, 1999 are dimes

will continue to correctly describe the contents of Clinton's pockets on this date under the above changes in background conditions as well as many others. If we want to distinguish those generalizations that intuitively are causal or explanatory, stability under changes in background conditions is not enough.<sup>5</sup>

A similar conclusion holds for the suggestion that what is crucial is stability under (some) changes in variables figuring in the relationship in question. The generalization (6.2.7) describing the correlation between the barometer reading  $B$  and occurrence of the storm  $S$  is stable under some changes in the values of  $B$  and  $S$ , for example, under changes in the values of  $B$  and  $S$  that result from

changing the value of  $A$ . However, this sort of stability is again not sufficient for the correlation (6.2.7) to represent a causal or explanatory relationship. As argued in chapter 2, the noncausal, nonexplanatory status of the correlation (6.2.7) shows itself in the fact that the correlation will disappear, break down, or fail to be invariant under any intervention that changes the value of  $B$ .

I conclude that it is crucial to the causal or explanatory status of a generalization that it would continue to hold under some interventions on the values of the variables figuring in the relationship. I emphasize this point because, as we shall see in more detail below, most other recent discussions of the relationship between lawfulness and stability, including those of Skyrms and Mitchell, assign no particular importance to invariance under interventions as opposed to stability under other sorts of changes.

What exactly do we mean by invariance under interventions? As motivation for the remarks that follow, consider a variation on an example from chapter 2. Suppose that a light is attached to a switch, and consider the generalization (6.2.8), which claims that the light will remain off if the position of the switch is less than one radian (57 degrees) and will remain on for any switch position between one and two radians. We can represent this relationship by means of an equation: let  $L$  be a variable that takes the value 1 if the light is on, and 0 otherwise; let  $q$  be the angular displacement of the switch measured in radians; and let  $[−]$  be the whole part function, so that, for example  $[1.57079 \dots] = 1$ . Then the relationship between  $L$  and  $q$  is:

$$(6.2.8) \quad L = [q].$$

Now consider an intervention that changes the position of the switch from 0 to 30 degrees. If the light remains off, there is an obvious sense in which the generalization (6.2.8) “continues to hold” or is “stable” under this intervention. Intuitively, however, this is not enough to establish that (6.2.8) is invariant, at least if we are trying to formulate an invariance condition that is sufficient for (6.2.8) to figure in explanations or to describe a causal relationship. The problem is that (6.2.8)’s continuing to hold under this intervention is compatible with the switch being broken, so that the light would not go on even if the switch were moved to some position between one and two radians. If this were the case, the position of the switch at 30 degrees would not cause or explain the light’s being off. The characterization of causation in chapter 2 reflects this: we required there that if  $q$  is to cause  $L$ , there must be an intervention on  $q$  that would change the value of  $L$ . If the switch is broken, no intervention on  $q$  will change the value of  $L$ . Similarly, the what-if-things-had-been-different condition on explanation formulated in chapter 5 requires that if the position of the switch is to explain why the light is off, it must be the case that there is some change in the position under which the value of  $L$  would have been different, that is, some switch position under which  $L = 1$ . Again, this condition is not met if the switch is broken.

This example suggests that the requirement we want to capture (if we want to formulate a connection among causation, explanation, and invariance) is

not just the idea that a generalization should continue to hold under some intervention on its independent variables, but rather that the generalization should continue to hold under an intervention that changes its independent variables sufficiently (or in such a way) that the value of its dependent variable is predicted by the generalization to change under the intervention. Thus, what matters for the invariance of (6.2.8) is whether (6.2.8) would continue to hold under (some) interventions that change  $q$  to a value greater than one radian, that is, whether it is really true, as (6.2.8) claims, that the light would go on under this intervention.

More generally, consider a change-relating generalization  $G$  relating a variable  $X$  to a second variable  $Y$ , and suppose that  $G$  holds, in the sense that it correctly describes the pattern of correlation between  $X$  and  $Y$ , for some particular system  $S$ . Consider an intervention (meeting the conditions **IN** in chapter 3) that changes the value of  $X$ , say  $x_0$ , that presently holds in  $S$  to some other value  $x_1$ , where  $x_1$  is claimed by  $G$  to be associated with a value of  $Y$  that is different from the value associated with  $x_0$ . That is,  $x_0 \neq x_1$  and  $G(x_0) = y_0 \neq G(x_1) = y_1$ . Following Woodward and Hitchcock (2003), let us call such interventions *testing interventions*. Such interventions do not merely change the value of  $X$  but change it in such a way that, according to  $G$ , the value of  $Y$  will change under the intervention. In this sense, such interventions “test”  $G$ .  $G$  is invariant under this testing intervention if and only if it correctly describes what the new value of  $Y$ ,  $y_1$ , would be under this change; that is, if and only if it remains true that  $G(x_1) = y_1$  for the system  $S$ . Invariance under at least one testing intervention (on variables figuring in the generalization) is necessary and sufficient for a generalization to represent a causal relationship or to figure in explanations. I emphasize that, as with other interventions, a testing intervention involves an actual physical change in the value of  $X$  for the system  $S$ . Also, as before, what matters is not whether the intervention is or even could be carried out, but whether  $G$  would continue to hold if it *were* to be carried out. Invariance is thus a modal or counterfactual notion.

In what follows, rather than using the cumbersome phrase “invariant under some testing interventions on variables figuring in a generalization,” I instead often use the simpler expression “invariant under interventions” or even “invariant.” For those readers who find the contrast between “interventions” and “testing interventions” confusing or distracting, it will do no harm to simply add another clause to the characterization of interventions **IN** in chapter 3, so that to qualify as an intervention, a change must meet the conditions in that characterization *and* be a testing intervention.

Let us see how this characterization applies to some specific examples. Consider a sample of gas  $X$  that conforms to the generalization

$$(6.2.9) \quad PV = nRT$$

in the sense that this generalization accurately describes the pattern of correlation or variation in the values of these variables in  $X$  (up to some appropriate level of approximation). Suppose that  $X$  is confined to a fixed

volume  $V = v$  within a rigid container, and consider an intervention that increases the temperature of the gas from  $T = t_0$  to  $T = t_1$ . (6.2.9) predicts that under this intervention, the pressure of the gas will change from  $p_0 = nRt_0/v$  to a different value  $p_1 = nRt_1/v$ , and so the intervention is a testing intervention. Whether (6.2.9) is invariant under this intervention will depend on whether this prediction is correct: whether (6.2.9) continues to accurately describe (again up to some suitable level of approximation) the relationship between  $P$ ,  $V$ , and  $T$  when this intervention occurs. For some samples of gas, it presumably is true that (6.2.9) is invariant under some testing interventions on  $T$ .

This example illustrates two other features of invariance that deserve emphasis. First, although for some systems, the ideal gas law (6.2.9) is invariant under some interventions on  $T$ , it is presumably not true that there are any systems for which (6.2.9) is invariant under all possible interventions on  $T$ . Instead, for any sample of gas, (6.2.9) will be invariant only for a certain range of (interventions that set) temperature values. If we increase the temperature of the gas sufficiently, intermolecular forces become important and the behavior of the gas will no longer conform to (6.2.9), even approximately. This illustrates the point made more abstractly earlier, that a generalization can be invariant under some interventions but not others. Second, interventions always occur within a particular system: they are changes in the value of some variable that characterizes that system, such as the temperature of a particular sample of gas. Hence, when we talk about invariance, this should be understood as having to do with stability of a generalization under interventions occurring in some particular system. Obviously, even if (6.2.9) is invariant under interventions that change the temperature of the particular sample of gas  $X$  from  $t_0$  to  $t_1$ , it does not follow—indeed it is false that—it will be invariant under such an intervention on temperature for all other systems  $X'$ . For example, (6.2.9) will presumably fail to be invariant—indeed, it will completely fail to characterize the behavior of—a sample of liquid under interventions on its temperature.

Both of these observations have implications for the explanatory import of invariant generalizations. First, as we shall see below, a generalization can be explanatory with respect to some aspects of the behavior of a system as long as it is invariant under some (appropriate) interventions even if it is not invariant under all interventions. For example, to explain why the pressure of the gas in the above example has increased from  $p_0$  to  $p_1$ , we can cite the generalization (6.2.9) and the fact that the temperature has increased from  $t_0$  to  $t_1$  as long as (6.2.9) would be invariant under an intervention that brings about this temperature increase, even though, as we have seen, (6.2.9) is not invariant under some other interventions that change temperature, and even though we may be unable to describe in any very exact way the full range of interventions over which (6.2.9) is and is not invariant. Second, what matters for explanation is invariance of a generalization with respect to the particular system  $X$  whose behavior we are trying to explain. In particular, as long as (6.2.9) is invariant under interventions that occur in  $X$ , we may appeal to it to

explain aspects of the behavior of  $X$  even though there are other systems for which (6.2.9) fails to hold and even though we cannot characterize in any very precise way the difference between  $X$  and these other systems.

The generalization (6.2.9) is invariant under some testing interventions and hence explanatory. We may compare (6.2.9) with the generalization (6.2.7) describing the correlation between  $B$  (barometer reading) and  $S$  (occurrence/nonoccurrence of a storm). Although this generalization is change-relating, it is not invariant under any testing interventions on  $B$ . If we change the value of  $B$  sufficiently that, according to (6.2.7),  $S$  should change under this intervention, we will find that the previously existing correlation between  $B$  and  $S$  breaks down. Thus, we cannot appeal to this generalization and the barometer reading to explain the occurrence or nonoccurrence of the storm.

As a second illustration, consider again the generalization (6.2.6), which claims that all of the coins in Clinton's pocket on a certain date are dimes. If (6.2.6) is to qualify as invariant at all, it must be interpretable as a change-relating generalization. If we take (6.2.6) literally, it is not a change-relating generalization: it says nothing about the denominations of coins that are not in Clinton's pocket. Hence, the notion of invariance under interventions does not apply and (6.2.6) is not explanatory. This agrees with intuition: when interpreted in this non-change-relating way, (6.2.6) does not even purport to tell us what a coin's denomination depends on.

Suppose instead that we do interpret (6.2.6) as a change-relating generalization, as claiming that, whereas all coins in Clinton's pocket are dimes, some coins not in Clinton's pocket are nondimes, so that whether or not a coin is located in Clinton's pocket ( $X$ ) is correlated with whether or not it is a dime ( $Y$ ). A testing intervention in this case will involve a change in the value of  $X$ —a change in the location of a coin—that, according to (6.2.6), is associated with a change in the value of  $Y$ . For example, introducing a dime into Clinton's pocket will not qualify as a testing intervention, because (6.2.6) predicts that under this change there will be no change in the value of  $Y$ , that is, no change in the denomination of the coin. The fact that (6.2.6) would continue to hold under this intervention does not establish that it is, in the relevant sense, invariant. On the other hand, the introduction of a penny (previously outside of Clinton's pocket) into his pocket does qualify as a testing intervention because (6.2.6), interpreted as a change-relating generalization, does predict that the value of  $Y$  would be different under this intervention. However, (6.2.6) is plainly not invariant under such interventions: the introduction of nondimes into Clinton's pocket will not transform them into dimes. Similarly, on the change-relating interpretation of (6.2.6) adopted above, the withdrawal of a dime from Clinton's pocket will also count as a testing intervention, because according to this interpretation of (6.2.6), coins that are outside of Clinton's pocket are more likely to be nondimes. Again, however, (6.2.6) is not invariant under this intervention, because withdrawing a coin from Clinton's pocket has no tendency at all to change its denomination. In fact, (6.2.6), interpreted as change-relating, is not invariant under any testing interventions on the variable  $X$  and hence cannot be used to

explain why, say, some particular coin that is in Clinton's pocket is a dime. Again, this agrees with preanalytic judgment.

Whether or not a generalization that relates changes is invariant under testing interventions on variables figuring in the relationship is closely connected to whether it describes a relationship that is hypothetically exploitable for purposes of manipulation and control—hypothetically exploitable in the sense that, although it may not always be possible, as a practical matter, to intervene to change the values of the quantities described by the variables that figure in the generalization, we can nonetheless think of the generalization as telling us that if it were possible to change those values, one could use them to change others. Thus, because the ideal gas law (6.2.9)  $PV = nRT$  is invariant under a range of interventions that change temperature, it correctly describes how, by manipulating the temperature of a gas and holding its volume constant, one could change its pressure. Similarly, because the gravitational inverse square law is invariant under a range of testing interventions that change mass and distance, we may think of it as telling us how, if we or some natural process were to manipulate these quantities in some system of gravitating masses, the gravitational force they exert would change in a systematic way. By contrast, because the barometer/storm correlation (6.2.7) and the generalization about the coins in Clinton's pocket (6.2.6) are not invariant under testing interventions on the variables that figure in them, they do not describe relationships that are hypothetically exploitable for purposes of manipulation and control. One cannot manipulate whether a storm occurs by fiddling with a barometer dial, and one cannot manipulate the denomination of a coin by moving it into or out of Clinton's pocket. If the argument of previous chapters is correct, this contrast is closely bound up with the fact that the inverse square law and (6.2.9) describe a causal or explanatory relationships, whereas (6.2.6) and (6.2.7) do not.

My argument so far has been that a necessary and sufficient condition for a generalization to represent a causal or explanatory relationship is that it be invariant under some testing interventions on variables occurring in the relationship. However, it is not part of my position that when a generalization represents a causal or explanatory relationship, it will continue to hold *only* under interventions. Typically, when a generalization describes a causal relationship, it will be stable under many other changes as well, including changes in both of the two categories described above: changes in background conditions and changes in values of the variables occurring in the generalization that do not involve interventions. This is true, for example, of the gravitational inverse square law (6.1.3), as we have already noted. Although in my view, a relationship can count as causal or explanatory if it fails to be invariant under some changes in background conditions, it seems clear that relationships that are, so to speak, almost endlessly sensitive to changes in background conditions—that are altered or disrupted in indefinitely many ways as conditions not included in the generalization shift—are regarded as of little scientific interest, even if they may be invariant under interventions in some highly specialized background circumstances. A similar conclusion seems

to hold regarding the significance of failures of stability under changes in variables figuring in a relationship that do not involve interventions. It is assumed in virtually all areas of scientific investigation that causal factors will at least sometimes behave in the same way in naturally occurring contexts that do not involve interventions as they do when changed by interventions; indeed, there would be little point to doing experiments if this were not the case. As Nancy Cartwright and Martin Jones (1991) put it in a recent paper, one makes the "standard assumption" that "the strength of C's . . . propensity to produce *E* is the same whether *C* occurs naturally or is caused by *d* [where *d* is a causal process having roughly the characteristics of an intervention]" (p. 221).

It may be tempting to conclude from these observations that it is only a necessary and not a sufficient condition for a generalization to describe a causal relationship that it be invariant under some testing interventions and that an additional necessary condition (or conditions) is (are) required: namely, that the generalization be stable under some range of changes in background conditions and/or non-*I*-changes. My view is that once we have decided to relativize the notion of invariance, this additional move is unnecessary. Rather than struggling with the problem of finding a nonarbitrary answer to the question of which background conditions or which non-*I*-changes a generalization must be stable under to qualify as invariant, it seems preferable and less arbitrary to say simply that different causal claims will differ in the range of changes in background conditions and non-*I*-changes under which they are invariant, and that we can spell out the content of different causal claims by being explicit about the range of interventions and other changes over which they are invariant. On this view of the matter, to qualify as invariant at all a generalization must be stable under some testing interventions (this constitutes the "threshold" for invariance), but among generalizations meeting this condition, there will be differences in range of invariance, represented by differences in the changes in background conditions and non-*I*-changes over which the generalization would continue to hold.

My concern so far has been to formulate a notion of invariance that is relevant to change-relating generalizations and hence to successful explanation. Is there a notion of invariance that is applicable to non-change-relating generalizations? As suggested above, I believe that there is. Although the notion of a testing intervention is not applicable to non-change-relating generalizations, we may still ask, concerning such a generalization, whether it will continue to hold under other changes in background conditions or for different values of the variables figuring in the generalization. To the extent that this is so, we may think of it as a relatively invariant generalization in a sense of "invariant" appropriate to non-change-relating generalizations. Thus, although the generalization "No material object can be accelerated from a velocity less than that of light to a velocity greater than light" is not a change-relating generalization, it is nonetheless true that there are no possible changes, nothing that we or nature can do, that will render it false. In this sense, it is invariant. Moreover, in my view, it is because it is invariant in this sense that we think of it as a law.

### 6.3 Which Relationships Are Invariant?

I said above that we expect that causal and explanatory relationships will be invariant both under some range of interventions and some non-I-changes. It is important, however, to understand just what this idea involves: it does not mean that we should always expect the overall observed association between cause and effect to be stable across both cases in which the cause occurs as a result of an intervention and cases in which it occurs as a result of a non-I-change. It may be that although there is a relationship between  $X$  and  $Y$  that is stable, what it stable is not the overall correlation between  $X$  and  $Y$ .

Consider a simplified version of an example drawn from Daniel Hausman (1998). A salt solution flows through pipes and mixing chambers. The concentration of salt in each chamber, measured by the variables  $X_1 \dots X_4$ , depends on the upstream chambers to which it is connected, according to the following equations:

$$(6.3.1) \quad X_2 = aX_1$$

$$(6.3.2) \quad X_3 = bX_1$$

$$(6.3.3) \quad X_4 = cX_2 + dX_3$$

Suppose first that the value of  $X_2$  is increased by one unit by means of an intervention, and consider the resulting change in the value of  $X_4$ . Such an intervention sets the value of  $X_2$  to its new level in a way that is independent of other causes of  $X_4$  like  $X_1$  that affect  $X_4$  via a route that does not go through  $X_2$ . In effect, we imagine an additional exogenous causal arrow or pipeline directed into  $X_2$  (besides the  $X_1 \rightarrow X_2$  pipe), with the change in level of  $X_2$  being due to this new route. If (6.3.1)–(6.3.3) correctly represents the causal structure of this system, then under this intervention the level of  $X_4$  should increase by  $c$  units.

Now suppose, by way of contrast, that  $X_2$  again increases by one unit, but this time the increase is the result of an increase in  $X_1$  by  $1/a$  units. This, of course, will not count as an intervention on  $X_2$  with respect to  $X_4$ , because  $X_1$  affects  $X_4$  via a route that does not go through  $X_2$ . The increase in  $X_1$  will produce an increase of  $b/a$  units in  $X_3$ , and this in turn will produce an additional change of  $(b/a)d$  units in  $X_4$ , along the  $X_1 \rightarrow X_3 \rightarrow X_4$  route. The total change in  $X_4$  will now be the contribution that  $X_2$  makes to  $X_4$  plus the contribution that goes through  $X_3$  or  $c + (b/a) \cdot d$  units.

Thus, the total change in  $X_4$  will be different for a unit change in  $X_2$  depending on whether  $X_2$  is changed by means of an intervention or instead is changed by changing  $X_1$ . Hausman takes this to show that it is a mistake to think that causal relationships should be expected to be invariant across both changes produced by interventions and changes that are not interventions (1998, pp. 222ff). Moreover, because he interprets the claim that the relationship between  $X_2$  and  $X_4$  is invariant as requiring that the total change in  $X_4$  that occurs when  $X_2$  is changed by a fixed amount must be the same



regardless of how  $X_2$  is changed, he takes this and similar examples to be clear counterexamples to the contention that there is a close link between causality and invariance.

In my view, this is an unreasonable conception of what invariance requires. The relationship that we should expect to be stable between  $X_2$  and  $X_4$  across different ways of changing  $X_2$  is not the observed overall association between  $X_2$  and  $X_4$ , but something more abstract: the functional relationship represented by equation (6.3.3). Assuming that the salt water apparatus operates in the way described, this relationship (6.3.3) is stable or invariant across different ways of changing  $X_2$ . Whether or not the value of  $X_2$  is changed by an intervention, it makes the same causal contribution to  $X_4$ : if the change in  $X_2$  is  $\Delta X_2$ , the change in  $X_4$  contributed by  $X_2$  is always  $c\Delta X_2$ , regardless of how that change in  $X_2$  is produced. Similarly,  $X_3$  makes the same contribution to  $X_4$ , as represented by the coefficient  $d$  in (6.3.3), regardless of how the change in  $X_3$  is produced. This is the respect in which  $X_2$  and  $X_3$  "act in the same way" across different contexts, including those that do and those that do not involve interventions. A true case of noninvariance would be, for example, a case in which (6.3.3) sometimes held and sometimes broke down, depending on how  $X_2$  was produced. The case under discussion does not have this character. (6.3.3), as well as (6.3.1) and (6.3.2), continue to hold when the same change in  $X_2$  is produced via an intervention and produced via a change in  $X_1$ . Indeed, Hausman himself assumes that this is the case when he uses (6.3.1)–(6.3.3) to calculate the very different total changes that occur in  $X_4$  when  $X_2$  is changed in different ways.

The same point holds for generalizations that are regarded as laws. Some writers (e.g., Giere 1999) construe the Newtonian gravitational inverse square law as a generalization connecting the total force experienced by some mass  $m_1$  to the quantities on the right-hand side of that law (a second mass  $m_2$ , and the distance between  $m_1$  and  $m_2$ ). On this construal, it follows that the law is false whenever other forces (gravitational or nongravitational) are incident on  $m_1$ , which is virtually always. It is far more plausible, and in accord with scientific practice, to construe the law as a claim relating the component of the gravitational force experienced by  $m_1$  which is due to the gravitational attraction of  $m_2$  to the quantities on the r.h.s. of the law. This relationship holds invariantly in a wide range of situations in which nongravitational forces are present.

A parallel point holds also for probabilistic causes. Michael Redhead (e.g., 1987) has argued that if  $k$  is a probabilistic cause of  $h$ , then  $p(h/k)$  should be "robust" in the sense that this probability is insensitive to (is invariant or does not change under) small changes in the way  $k$  comes about. Commenting on this proposal, Richard Healey (1992) writes:

In applying the condition of robustness to decide whether to count event  $k$  as cause of event  $h$ , it is important to take account of other causes of  $h$ . If an event  $h$  has two partial causes  $j$  and  $k$ , then one would not expect  $P(h/k)$  to remain invariant under modifications in the causal antecedents of  $k$  which do not leave  $j$  fixed. The effects of a partial cause often depend on what other partial causes act in concert with it. Plastic

explosives hidden in a cassette recorder would not have caused the destruction of Pan Am flight 103 over Lockerbie if the aircraft had not been flying at the time. But even a small modification in the causal antecedents of a partial cause of an event may alter other partial causes of that event. A false move by the terrorist who planted the explosives might have alerted Pan Am to the possibility that he had planted the device in one of their aircraft, prompting them to ground all their flights out of Europe just as flight 103 was about to leave London. (p. 286)

In this example,  $k$  causes  $h$ , but the conditional probability  $P(h/k)$  is not invariant under different possible ways of bringing about  $k$ . This does indeed show that the robustness or invariance of  $P(h/k)$  under different ways of bringing about  $k$  is not a necessary condition for  $k$  to cause  $h$ , but, as Healey recognizes, it does not rule out the possibility that there are other invariance requirements to which probabilistic causes conform (cf. chapter 7), and it does not undercut the more general claim that there is a connection between invariance and causality.<sup>6</sup>

#### 6.4 Degrees of Invariance

As already intimated, invariance is not an all-or-nothing matter. Most generalizations that are invariant under some interventions and changes in background conditions are not invariant under others. As we shall see shortly, we may legitimately speak of some generalizations as more invariant than others—more invariant in the sense that they are invariant under a larger or more important set of changes and interventions than other generalizations. Moreover, there is a connection between range of invariance and explanatory depth: generalizations that are invariant under a larger and more important set of changes often can be used to provide better explanations and are valued in science for just this reason. The picture that I defend is thus one in which there is both a threshold—some generalizations fail to qualify as invariant or explanatory at all because they are not invariant under any interventions on the variables that explicitly figure in the generalization—and above this threshold distinctions or gradations of various sorts in degree of invariance. This picture corresponds to how, intuitively, we seem to think about explanatory (or causal or nomological) relationships. Some relationships (e.g., the relationship described by the correlation between the barometer reading  $B$  and the occurrence of the storm  $S$ ) are not causal or explanatory at all, but among those that are, some may be used to provide deeper explanations than others. This represents just one of many points at which the invariance-based account I defend contrasts with more traditional frameworks for thinking about laws and their role in explanation. The traditional frameworks suggest a dichotomy: that either a generalization is a law or else it is purely accidental. Moreover, it is assumed that the boundary between laws and nonlaws coincides with the boundary between those generalizations that can be used to explain and those that cannot. The invariance-based account rejects both of these ideas.

The ideas introduced in the previous paragraph—that generalizations may differ in the range of changes or interventions over which they are invariant and that these differences are connected to differences in their explanatory status—are familiar themes in the econometrics literature. They are illustrated and endorsed by Tygre Haavelmo, one of the founding figures of econometrics, in a well-known passage from his monograph “The Probability Approach in Econometrics” (1944). In this passage, Haavelmo introduces a notion that he calls autonomy but that is really just another name for (one) kind of invariance condition. He writes:

If we should make a series of speed tests with an automobile, driving on a flat, dry road, we might be able to establish a very accurate functional relationship between the pressure on the gas throttle (or the distance of the gas pedal from the bottom of the car) and the corresponding maximum speed of the car. And the knowledge of this relationship might be sufficient to operate the car at a prescribed speed. But if a man did not know anything about automobiles, and he wanted to understand how they work, we should not advise him to spend time and effort in measuring a relationship like that. Why? Because (1) such a relation leaves the whole inner mechanism of a car in complete mystery, and (2) such a relation might break down at any time, as soon as there is some disorder or change in any working part of the car. We say that such a relation has very little autonomy, because its existence depends upon the simultaneous fulfillment of a great many other relations, some of which are of a transitory nature. On the other hand, the general laws of thermodynamics, the dynamics of function, etc., are highly autonomous relations with respect to the automobile mechanism, because these relations describe the functioning of some parts of the mechanism irrespective of what happens in some other parts. (pp. 27–28)

Haavelmo then suggests the following, more formal characterization of autonomy:

Suppose that it would be possible to define a class  $S$ , of *structures*, such that *one member or another* of this class would, approximately, describe economic reality in any practically conceivable situation. And suppose that we define some non-negative *measure* of the “size” (or the “importance” or “credibility”) of any subclass,  $W$  in  $S$  including itself, such that, if a subclass contains completely another subclass, the measure of the former is greater than, or at least equal to, that of the latter, and such that the measure of  $S$  is positive. Now consider a particular subclass (of  $S$ ), containing all those—and only those—structures that satisfy a particular relation “ $A$ .” Let  $W_A$  be this particular subclass. . . We then say that the relation “ $A$ ” is *autonomous* with respect to the subclass of structures  $W_A$ . And we say “ $A$ ” has a degree of autonomy which is the greater the larger the “size” of  $W_A$  as compared with that of  $S$ . (pp. 28–29; emphasis in original)

Although this characterization is far from completely transparent (among other things, Haavelmo does not tell us how to go about determining the “size” or “importance” of  $W$ , matters I address below), the underlying idea is perhaps

clear enough. In the most general sense, the degree of autonomy of a relationship has to do with whether it would remain stable or invariant under various possible changes. (As I have argued, if, like Haavelmo, we wish to use this idea to distinguish between those relationships to which we can appeal to [causally] explain and those that cannot be so used, we need to include, among the changes over which we demand that a relationship be autonomous, those that correspond to interventions on the variables figuring in the relationship.) The larger the class of changes under which the relation would remain invariant—the more structures in  $W$  compatible with relation—the greater its degree of autonomy. Haavelmo suggests that physical laws such as the laws of thermodynamics and fundamental engineering principles such as those governing the internal mechanism of the car will be relatively autonomous in this sense. By contrast, the relationship (6.4.1) between the pressure on the gas pedal and the speed of the car will be far less autonomous. We may imagine that (6.4.1) holds stably for some particular car if we intervene repeatedly to depress the pedal under sufficiently similar conditions and, if so, (6.4.1) will be invariant under some interventions. Nonetheless, (6.4.1) will be disrupted by all sort of changes: by variations in the incline along which the car travels, changes in the head wind, changes in the fuel mixture, changes in the internal structure of the car engine (e.g., by cleaning the spark plugs and adjusting the carburetor), and so on. (6.4.1) will also be disrupted by various “extreme” interventions on the gas pedal, for example, those that are sufficiently forceful that they destroy the pedal mechanism. (6.4.1) is thus relatively fragile or nonrobust in the sense that it holds only in certain very specific background conditions and for a restricted range of interventions. Intuitively, although (6.4.1) is invariant under some interventions and changes, it is invariant under a “smaller” set of interventions and changes than fundamental physical laws.

According to the account defended in chapter 5, if (6.4.1) holds invariantly for some range of testing interventions that depress the gas pedal by various amounts, for some type of car in a kind of environment, then we may appeal to (6.4.1) and to the depression of the pedal to explain the speed of the car, provided that the car is within the domain of invariance of (6.4.1). Within its domain of invariance, (6.4.1) describes a relationship that can be exploited for purposes of manipulation and control: it describes how we can change the speed of the car by changing the depression of the gas pedal. This is a feature that (6.4.1) shares with paradigmatic laws like the gravitational inverse square law and that distinguishes both from purely accidental generalizations like that describing the correlation  $B$  and  $S$ . Because (6.4.1) does not completely lack invariance, an explanation appealing to (6.4.1) will exhibit, albeit in a very limited way, the pattern of counterfactual dependence (involving counterfactuals about what would happen under interventions) that, according to chapter 5, is at the heart of successful explanation. We can appeal to (6.4.1) to explain even if, because of its relative fragility or for other reasons, we are unwilling to regard it as a law of nature. We can thus think of this example as illustrating my claim that it is invariance and not lawfulness per se that is crucial in explanation.

However, like Haavelmo, I also take it to be obvious that an explanation of the speed of the car that appeals just to (6.4.1) is shallow and unilluminating. I follow Haavelmo in tracing this to the fact that the relation (6.4.1) is relatively fragile: it is invariant only over a very limited range of interventions and changes in background conditions, and can be used to answer only a very limited range of what-if-things-had-been-different questions. A deeper explanation for the behavior of the car would need to appeal to (6.4.2) laws and engineering principles, like those mentioned by Haavelmo, that are invariant under a much wider range of changes and interventions. Not coincidentally, such a deeper explanation could be used to answer a much wider range of *w-questions*. For example, unlike (6.4.1), the generalizations (6.4.2) appealed to in this deeper explanation are such that they could be used to explain why the car moves with the speed that it does over a variety of different kinds of terrain and road conditions, under a variety of different kinds of mechanical changes in the internal structure of the car, and so on. The what-if-things-had-been-different account of explanation thus seems to capture the relevant features of Haavelmo's example in a very natural way.

What might it mean to say, as Haavelmo does, that one generalization is invariant under a "larger" set of changes or interventions than another? In Haavelmo's example, this question has a straightforward answer. To a very good degree of approximation, the range of changes and interventions over which (6.4.1) is invariant is a proper subset of the range of changes and interventions over which the generalizations (6.4.2) of the deeper engineering theory of the behavior of the car are invariant. That is, any change that will disrupt the latter will also disrupt (6.4.1), but not vice versa. Thus, any properly behaved measure will assign a larger size to the domain of invariance of the latter.

A similar basis for comparison exists in the case of many other pairs of generalizations, for example, the ideal gas law  $PV = nRT$  and the van der Waals force law

$$(6.4.3) \quad [P + a/V^2][V - b] = RT$$

Here,  $a$  and  $b$  are constants characteristic of each gas, with  $b$  depending on the diameter of the gas molecules and  $a$  on the long-range attractive forces operating between them. For any given gas, the generalization (6.4.3) holds invariantly in circumstances in which the ideal gas law holds, but it also holds invariantly in at least some circumstances—roughly those in which intermolecular attractive forces are important and in which the volume of the constituent molecules of gas are large in comparison with the volume of the gas—in which the ideal gas law breaks down. The range of changes or interventions over which the van der Waals force law (6.4.3) is invariant is again "larger" than the range of changes over which the ideal gas law is invariant in the straightforward sense that the latter set of changes is a proper subset of the former. Moreover, just as in Haavelmo's example, this larger range of invariance means that we can use (6.4.3) to answer a larger set of what-if-things-had-been-different questions than the ideal gas law. Thus, we

can use (6.4.3) to answer questions not just about what would happen to the values of one of the variables  $P$ ,  $V$ , and  $T$ , given changes in the others in circumstances in which intermolecular forces are unimportant and intermolecular distances are large in comparison with molecular volumes, but also about what would happen to  $P$ ,  $V$ , or  $T$  when these conditions no longer hold. We can also use the van der Waals equation to explain various phenomena having to do with phase transitions: again, circumstances in which the simpler ideal gas law breaks down. A similar relationship holds between many other pairs of generalizations, for example, between the laws of General Relativity and those of Newtonian gravitational theory.

It is important to understand that the claim that the range of interventions over which generalization ( $G_1$ ) is invariant is a proper subset of the range of changes and interventions over which a second generalization ( $G_2$ ) is invariant is *not* merely a restatement of the claim that ( $G_1$ ) and ( $G_2$ ) are both true and that ( $G_1$ ) is derivable from ( $G_2$ ) but not vice versa. For one thing, ( $G_1$ ) may be derivable from ( $G_2$ ) but not vice versa even if neither generalization is invariant at all. For example, the true generalization ( $G_1$ ) that all spatiotemporal regions of 1 meter radius within 10 light years of the earth contain cosmic background radiation at approximately 2.7 degrees K is derivable from the generalization ( $G_2$ ) that all spatiotemporal regions in the universe contain such background radiation, but neither generalization is invariant—neither is change-relating—and both may well depend in an extremely sensitive way on the initial conditions obtaining in the early universe (see section 6.12).

As another illustration, consider the conjunction of Galileo's law (construed as a generalization in which the height  $h$  above the surface of the earth from which an object is dropped is the independent variable and time  $t$  of fall the dependent variable) and the ideal gas law. There is an obvious sense in which this conjunction is "more general" than either conjunct taken separately: the conjunction makes interesting predictions about both falling objects and gases, whereas each conjunct taken separately makes predictions about only one of these classes of objects. In the language of section 6.6, the conjunctive generalization has greater "scope" than each of its conjuncts. Yet it seems clear that if we want to explain why an object took three seconds to fall, an explanation that cites the conjunctive generalization is no deeper or more illuminating than the one that cites Galileo's law alone. The account of explanation and invariance I have been defending supports this judgment: the conjunctive law provides no more information about what will happen under interventions on those variables (in this case, the variable  $h$ ) affecting the falling time of objects in free fall than does Galileo's law alone. By the same token, the conjunctive generalization is not invariant under a wider range of interventions on the variable  $h$ , which figures in the explanation of falling time, than Galileo's law alone. This corresponds to our assessment that the conjunction tells us nothing more than Galileo's law about the factors on which falling time of fall depends.

Galileo's law is invariant under some interventions on the height from which the object is dropped, but it would fail to hold if the object were dropped from a height that is large in relation to the earth's radius or if it were dropped

from the surface of a body with a mass very different from that of earth. Newton's second law together with his law of gravitation entail a generalization  $G'$  that also allows us to compute the time it would take an object to fall a certain distance.  $G'$ , however, is not restricted in the way that Galileo's law is: it will remain invariant under changes in the mass and radius of the massive body on which the object is dropped. It thus has a greater range of invariance than Galileo's law, achieving this by explicitly incorporating the mass and radius of the planet (or whatever) as variables. The greater range of invariance of  $G'$  in comparison with Galileo's law in turn is connected to our judgment that  $G'$  provides a deeper explanation of why objects dropped from various heights take the time to fall that they do—that it does a better job than Galileo's law at identifying the conditions on which time of fall depends. The larger point that is illustrated by these examples is that when we compare generalizations with respect to range of invariance, we are comparing them along a very specific dimension of generality. Such comparisons are very different from the comparisons that we make when we simply ask whether one generalization is derivable from another.<sup>7</sup>

Although we may compare the range over which two generalizations are invariant when the proper subset relation just described holds, this obviously yields only a partial ordering. For many pairs of generalizations, neither will have a range of invariance that is a proper subset of the other. Moreover, the proper subset relation provides at best a basis for ordinal comparisons. We can say that one generalization is invariant under a larger set of changes than another, but we have no basis for claiming that this set is large or "important" (to use Haavelmo's word) in some more absolute sense. Is there some other basis on which we can make such claims? I believe there is. This basic idea is more easily illustrated than precisely characterized, but the underlying intuition is this: for different sorts of generalizations, applicable to different sets of phenomena or subject matters, there often will be specific sorts of changes that are privileged or particularly important or significant from the point of view of the assessment of invariance, privileged in the sense that it is thought to be especially desirable to construct generalizations that are invariant under such changes and that generalizations that are invariant under such changes are regarded as having a fundamental explanatory status in comparison with generalizations that are not so invariant. The privileged changes in question will be subject matter- or domain-specific: one set of changes will be important in fundamental physics, another in evolutionary biology, and yet another in microeconomics. Thus, expectations about the sorts of changes over which fundamental relationships will be invariant help to set the explanatory agenda for different scientific disciplines. These expectations will in turn be grounded in very general empirical discoveries about the sorts of relationships in the domains of these disciplines that have been found to be invariant in the past and under what sorts of changes.

As an illustration, consider the symmetry requirements that fundamental physical laws are expected to satisfy. These requirements are rooted in very general empirical facts about the natural world: that relationships can be

found that are invariant under certain kinds of changes (such as changes from one inertial frame to another) and not other changes is an empirical discovery. These empirical discoveries in turn generate expectations about the kinds of symmetries physical laws should exhibit. At present, generalizations that fail to satisfy such symmetry requirements are unlikely to be regarded as candidates for fundamental laws or explanatory principles, regardless of whatever other features they possess. The requirements thus have a special or privileged status: from the point of view of the assessment of invariance in physics, they are more important than invariance under other sorts of changes.

For purposes of comparison, consider what counts as an important kind of change for the purposes of assessing invariance in contemporary microeconomics. In microeconomics, individual economic agents are often assumed to conform to the behavioral generalizations constituting rational choice theory (RCT). For present purposes, we may take these generalizations to include the principles of expected utility theory, as described, for example, in Luce and Raiffa (1957), together with the assumption that choices are self-interested in the sense that agents act so as to maximize quantities like personal income and wealth. Even if we assume, for the sake of argument, that these generalizations are roughly accurate descriptions of the behavior of many participants in markets, it is clear that there are many changes and interventions over which the generalizations will fail to be invariant. For example, there are many pharmaceutical interventions and surgically produced changes in brain structure that will lead previously selfish agents to act in non-self-interested ways or to violate such principles of RCT as preference transitivity. However, economists have not generally regarded these sorts of failures of invariance as interesting or important, at least if, as is often the case, they occur relatively rarely in the populations with which they deal.

By contrast, failures of invariance under other sorts of changes are regarded as much more important. For example, microeconomists often require that fundamental explanatory generalizations such as the principles of RCT be invariant under changes in information available to economic agents or under changes in their beliefs and under changes in the incentives or relative prices they face. Indeed, a standard assumption among many microeconomists—one might take it to be constitutive of a certain sort of methodological individualism—is that the generalizations that will be invariant under such changes in information and prices all describe the behavior of individual economic agents rather than the relations between macroeconomic or aggregate-level variables such as “inflation,” “unemployment,” and “gross domestic product.” That is, there are no purely macroeconomic relationships that are invariant under changes in information and incentives, and hence there are no fundamental explanatory relationships among macroeconomic variables.

As an illustration, consider the macroeconomic relationship known as the Phillips curve. This describes the historically observed inverse relationship or trade-off between unemployment and inflation in many Western countries from the mid-nineteenth to the mid-twentieth century.<sup>8</sup> A crucial question is whether this relationship is (or was) invariant under policy interventions on



these variables. According to some Keynesian models, the Phillips curve describes a relationship that is invariant under at least some governmental interventions that change the inflation rate. If so, governments would be able, by increasing the inflation rate, to decrease the unemployment rate—a highly desirable result. The burden of an influential criticism of these models developed by Lucas (1983; the so-called Lucas critique) is that the relationship discovered by Phillips is not invariant under such interventions: the result of interventions that increase the inflation rate will not be to lower the unemployment rate but rather simply to produce changes in the Phillips curve itself. Very roughly, according to this critique, increasing inflation will reduce unemployment only if employers or employees mistake an absolute increase in prices for a favorable shift in relative prices and (given the assumption that these agents are “rational”) this is not a mistake they will make systematically or for any length of time. As soon as these agents realize that a general increase in the price level has occurred or come to expect that such an increase will occur, unemployment will return to its original level. To put the point abstractly, the Phillips curve is not invariant under changes in the information available to economic agents or under changes in their expectations of a sort that almost certainly will occur once the government begins to intervene to change the inflation rate. A similar point will hold for many other macroeconomic relationships.

This example illustrates how issues about invariance arise naturally in economics. The interesting question for economists is not whether the Phillips curve is a law of nature or completely exceptionless, but whether it is invariant under certain specific kinds of changes and interventions. If the Phillips curve is not invariant under the relevant sorts of interventions, it will not be regarded as a fundamental economic relationship or as a relationship that it would be satisfactory to take as primitive in a deep economic explanation. (For example, if the Lucas critique is correct, it would be unsatisfactory to appeal to the inflation rate and the Phillips curve to explain the unemployment rate.) This is not because it fails to be invariant under all possible changes and interventions (including all-out nuclear war or radical psychosurgery performed on the entire U.S. population), but because it (allegedly) fails to be invariant under a specific set of possible changes that are thought to be particularly important: changes in the information that economic agents receive.

My suggestion, then, is that both of the considerations described in this section—comparisons of invariance based on the proper subset relation and judgments about the significance or importance of the intervention over which a generalization is invariant—play an important role in the construction and assessment of explanatory generalizations. Together, they provide a partial basis for distinguishing among invariant generalizations with respect to degree and kind of invariance and for judging that, although a generalization is invariant under some interventions, it is nonetheless relatively fragile or unrobust in the sense that it is stable only under an unimportant set of interventions or under a set of changes that is relatively small in comparison with some rival generalization. As remarked above, the idea that generalizations

can differ in the range or importance of the interventions over which they are invariant is one of a number of respects in which the invariance-based framework I recommend departs from the traditional law-based framework. In contrast to the traditional framework, which admits just two mutually exclusive possibilities (a generalization is either a law or else it is “accidental”), the notion of invariance permits a much richer set of distinctions among invariant generalizations. As we shall see, this makes the invariance-based framework much better suited for capturing the characteristics of explanatory generalizations in the special sciences.

I conclude this section with a caveat: both whether one generalization is “more invariant” than another and “explanatory depth” are complicated and multidimensional notions. The remarks in this section are intended to describe *some* considerations that are relevant to assessing degree of invariance or explanatory depth but they are emphatically *not* intended to be exhaustive and comprehensive. As an illustration of some relevant additional complexities, consider the following example. Galileo’s law of free fall ( $G$ ) relates the amount of time it takes an object to fall to earth to the height from which it was dropped:  $H = aT^2$ . Compare ( $G$ ) with the following ‘Goodmanized’ version of Galileo’s law ( $G^*$ ):  $H + (H - h_0)(H - h_1) \dots (H - h_n) = aT^2$ . Suppose that the object whose time of fall we wish to explain was in fact dropped from a height of  $h_0$ . Then this new relationship will be invariant under some testing interventions—those that change the value of  $H$  from  $h_0$  to  $h_1 \dots, h_n$ . In this case, ( $G^*$ ) strikes us as at best very weakly explanatory, if indeed it is explanatory at all. Presumably the problem with ( $G^*$ ) lies in the disjoint nature of the set of values of  $H$  for which it holds. In order to get to a testing intervention under which ( $G^*$ ) is invariant, we must “skip over” a much larger set of testing interventions under which ( $G^*$ ) is not invariant. If the true functional relationship between  $X$  and  $Y$  is  $Y = F(X)$  any other sufficiently wildly oscillating function  $Y = H(X)$  relating  $X$  to  $Y$  will hit the right values of  $Y$  for some values of  $X$ , but  $H$  will have very dubious explanatory credentials, especially if  $F$  is available.

A natural way of handling examples of this sort would be to require that at least in the case of relationships among variable that take continuous values, an explanatory relationship should be invariant under all testing interventions that change the values of variables to new values within some neighborhood of the actual values. ( $G^*$ ) fails to conform to this requirement because any neighborhood of  $H = h_0$  will contain values of  $H$  such that ( $G^*$ ) is not preserved when  $H$  is set to those values by interventions. More generally, it is clearly relevant to the explanatory credentials of a generalization that the range of values over which it is invariant not be too disjoint or discontinuous.<sup>9</sup>

## 6.5 Invariance and the Traditional Criteria for Lawfulness

Philosophers have traditionally employed a number of criteria to distinguish between laws and accidental generalizations. Laws are said to be exceptionless

generalizations representable by universally quantified conditionals, to contain only purely qualitative predicates and/or natural kind terms and to make no reference to particular objects or spatiotemporal locations, to have very wide scope, to support counterfactuals, to be projectable or confirmable by their instances, to be integrated or potentially integrable into a body of systematic theory, and to play a unifying or systematizing role in inquiry. To have a useful shorthand way of referring to these criteria, let us call them the traditional criteria for nomological status.

My view is that most of these criteria are not helpful either for understanding what is distinctive about laws or for understanding the features that characterize explanatory generalizations, either in the special sciences or in disciplines like physics and chemistry. In general, it is the range of interventions and other changes over which a generalization is invariant and not the traditional criteria that are crucial both to whether it is a law and to its explanatory status. Moreover, whether a generalization is invariant (and if so, over what range of changes and interventions) is surprisingly independent of most of the traditional criteria, a point that is perhaps suggested by the fact that in appealing to the notion of invariance to describe the differences among various generalizations described above (the correlation between *B* and *S*, the ideal gas law, etc.) we did not need to explicitly invoke these traditional criteria. In fact, a generalization may satisfy many of the traditional criteria and yet fail to be invariant, and a generalization may be invariant even though it fails to meet many of the traditional criteria. Among the traditional criteria, only one (support for counterfactuals) is relevant to whether a generalization counts as invariant, and even then, as we shall see in section 6.9, this criterion is understood quite differently on the invariance-based approach than on the traditional approach. Seen from the perspective of the invariance-based approach, the fuzziness and inadequacy of the traditional criteria for lawfulness are thus neither surprising nor alarming. If what we are interested in is which generalizations can figure in explanations, we don't have to address the question of whether such generalizations are "laws" in the sense of satisfying the traditional criteria. Instead, we can simply bypass these criteria and focus directly on the question of whether the generalizations of interest are invariant in the right way.

There are a number of different ways of responding to these observations. One is to drop the concept of a law of nature as unhelpful for understanding science and to focus directly on the notion of invariance, since the latter notion captures, or so I have suggested, what is really relevant to successful explanation. This strategy fits, at least in some respects, the views of those philosophers (Cartwright 1983b; Giere 1988; van Fraassen 1989) who, on various grounds, have been skeptical about the roles standardly assigned to laws in science. (Of course, these philosophers may not find my positive views about the role of invariance in explanation congenial.) As indicated earlier, I do not adopt this sort of wholesale skepticism about the notion of law. Instead, I proceed on the assumption that, although the notion is not especially sharp or clear (see section 6.10), generalizations like the gravitational inverse square

law, Maxwell's equations, and the field equations of General Relativity are indeed appropriately described as laws and that such laws are approximately true. What is confused or wrongheaded is not the notion of law as it actually figures in scientific practice in disciplines like physics and chemistry, but rather many of the criteria for (or theories about) lawhood proposed by philosophers.

A second possibility is to retain the notion of law for paradigms like Maxwell's equations but to regard the question of whether that notion should be extended to cover explanatory generalizations in the special sciences like Mendel's as a "purely verbal" or "don't care" issue. According to this position, it simply doesn't matter, independently of whether or not generalizations like Mendel's are invariant, whether we choose to regard them as genuine laws. We can, if we wish, stipulate that the word "law" must be used in such a way that all invariant generalizations are laws. If so, it will follow that the nomothetic thesis that all explanation requires laws is correct (because explanation requires invariance). It will also follow that, to the extent that they are invariant, Mendel's laws and other explanatory generalizations in the special sciences will qualify as laws. Alternatively, we may choose to regard similarity to paradigmatic laws as necessary for lawhood. If so, generalizations like Mendel's will probably not count as laws. Nonetheless, as long as such generalizations are invariant in the right way, they can figure in explanations. According to this "don't care" position, the difference between these two proposals about how permissively to use the word "law" is purely a matter of terminology and reflects nothing substantive.

Although I have considerable sympathy for this second position, I think, for reasons to be described in more detail in section 6.10, that it is not completely arbitrary how permissively we use the word "law." Instead, there are considerations that favor a relatively restricted notion of law, according to which generalizations like Mendel's or the regression equation like (5.6.1) relating editorial slant to vote difference do not count as laws. Fundamentally, this is because the more permissive usage conceals genuine differences between generalizations like Mendel's and (5.6.1), on the one hand, and generalizations like Maxwell's equations, on the other, and because there is no real motivation for adopting it, once it is recognized that invariance and ability to answer a range of *w-questions* are what matter in explanation. Thus, in what follows, rather than thinking of all invariant generalizations as laws, I urge instead that we think of laws as just one kind of invariant generalization. On this view of the matter, laws do indeed play an important role in (some areas) of science, such as physics, but they are both less central and less pervasive in science as a whole than traditional approaches suppose. Both in this respect and in my skepticism about whether there are many generalizations, even in physics, that conform to all of the traditional philosophical criteria for lawfulness, my views echo the doubts about the notion of law found in Cartwright, Giere, and van Fraassen. However, I differ from these writers in thinking that there is a defensible notion of law, different in important respects from the notion described by philosophers, that figures in disciplines such as physics.

## 6.6 Invariance, Qualitative Predicates, and Scope

I claimed above that whether a generalization is invariant is surprisingly independent of whether it satisfies most of the traditional criteria for lawfulness. Showing this in detail would require another book. In what follows, I illustrate this claim by focusing on just a few of these criteria: the requirements that laws must not refer to particular places or times and must not be too narrow in scope (this section), the requirement that laws and explanatory generalizations must be exceptionless (section 7), and the idea that laws must support counterfactuals, which I discuss and reinterpret in section 9. The criteria associated with the Mill–Ramsey–Lewis (MRL) theory of law are discussed in section 11.

Consider first the common suggestion that laws must contain “purely qualitative” predicates and must contain no essential reference to particular objects, times, or places. A number of writers (e.g., Nagel 1961, p. 57; Carroll 1994, pp. 36ff) have observed that this is a dubious criterion for lawfulness, both because the notion of a “purely qualitative” predicate is far from clear and because, to the extent that this notion is clear, there appear to be laws (like Galileo’s law of falling bodies) that do not respect the criterion. It is thus of considerable interest that it is perfectly possible for a generalization to be invariant without satisfying this criterion. To see this, imagine that Clinton’s pockets on January 8, 1999 do turn out to have the property that whenever a nondime is introduced into them by means of a testing intervention, it is turned into a dime. In such a case, the generalization

(6.6.1) All the coins in Clinton’s pockets on January 8, 1999 are dimes

interpreted in a change-relating way, would be invariant under such interventions despite the fact that it contains nonqualitative predicates and makes essential reference to a particular person and time. Indeed, we can imagine, consistently with the nonqualitative character of (6.6.1), that it is invariant under a very wide range of interventions and changes: that it continues to hold no matter how coins are introduced (and no matter which coins are introduced) into Clinton’s pockets and no matter which other background changes occur.

This particular example is, of course, fantastic, but others are less so. Aristotle is often represented as thinking that as a matter of law (6.6.2) all freely falling objects will move toward a particular spatial location: the center of the earth. He was, of course, wrong about this, but what was the nature of his mistake? If it is built into the very concept of law that laws must not refer to particular places, his mistake was a conceptual one. A much more plausible judgment is that his mistake was empirical. The connection between lawfulness and invariance for which I have been arguing supports this judgment, for it is clear that (6.6.2) might have turned out to be a highly invariant generalization despite its reference to a particular object or spatial location.

For similar reasons, it is perfectly possible for a generalization to be invariant only under changes and interventions that occur within a limited spatial or temporal interval and to break down outside that interval. Suppose that, contrary to actual fact, the Phillips curve turned out to be invariant under

governmental interventions that changed the inflation rate between, say, 1870 and 1970 in the United Kingdom, although not invariant outside this interval. If this had been the case, then (I would claim), despite the limited spatiotemporal scope of this relationship, one could appeal to it and to the fact that the U.K. government intervened to raise the inflation rate in 1915 to explain why unemployment fell after this intervention. More generally, in contrast to traditional law-based accounts of explanation, the notion of invariance allows us to talk about explanatory relations that hold only over limited spatiotemporal intervals or that make reference to particular objects, events, or processes. As we shall see below, many explanatory generalizations in the special sciences seem to have exactly these features, and this is one reason the notion of invariance is particularly well-suited to understanding their character.

Consider next the relationship between invariance and *scope*. This latter notion is difficult to characterize precisely, but I take the intuitive idea to be that a generalization has wide scope if it holds for a “large” range<sup>10</sup> of different systems or different kinds of systems, in the sense that the systems in question satisfy both its antecedent and its consequent and that we can appeal to the generalization to describe the behavior of such systems. We think of the Newtonian inverse square law as having wide scope because it holds for all masses throughout the universe: for bodies falling near the surface of the earth, for all planets orbiting the sun, and so on. By contrast, we think of a version of Hooke’s law

$$(6.6.3) \quad F = -K_s x$$

describing the behavior of one particular sort of spring  $S$ , characterized by the specific spring constant  $K_s$ , as much narrower in scope. Most other sorts of springs will obey a different (or no) version of Hooke’s law, and most systems that are not springs will not be describable in terms of this law at all.

Although we are reluctant to regard generalizations with narrow scope as laws (see section 6.10), it is nonetheless of considerable interest that the scope of a generalization seems to have little to do with whether it is invariant and, if so, over which changes and hence, on my view, little to do with explanatory import. Despite its narrowness of scope, the generalization (6.6.3) might well turn out to be invariant under a substantial range of interventions that change the extension of a particular spring or set of springs and under other changes as well. If so, according to the account of explanation advocated here, we can appeal to (6.6.3) to explain why a particular spring exerts the force it does. Again, this is important for understanding explanation in the special sciences. Many generalizations in the special sciences, such as the regression equations described in chapter 7 and generalizations about particular biological mechanisms, lack broad scope—intuitively, they are about very specialized kinds of systems—but it would be a mistake to conclude on this ground alone that they are unexplanatory.

Because the distinction between invariance and scope will be important subsequently, it is worth spending a little more time to make it clear. Intuitively,

the scope of (6.6.3) has to do with how many different kinds of springs (or perhaps, alternatively, with how many individual springs) are correctly described by (6.6.3). The scope of (6.6.3) would be greater if it correctly described the behavior not only of springs made out of the material that is characteristic of *S* but also other sorts of springs, found elsewhere in the universe and made of very different kinds of material. The scope of (6.6.3) would be narrow if there were only one spring in the universe conforming to (6.6.3) and greater if such springs were very common. By contrast, when we ask about the range of invariance of (6.6.3) qua description of the behavior of *S*, we asking a different kind of question. In this case, we want to know, for springs of this very type (regardless of how many there are), the range of interventions that change the extension of the spring and the range of changes in background conditions over which (6.6.3) is invariant. It doesn't follow merely from the fact that (6.6.3) has wide scope that it is invariant under a wide range of interventions and changes: (6.6.3) might correctly characterize the behavior of lots of different kinds of springs, but it might be the case that for each of these kinds of springs, (6.6.3) is invariant only under a narrow range of interventions. Conversely, even if there was only one spring in the universe conforming to (6.6.3), (6.6.3) could still be a highly invariant generalization concerning the behavior of that spring.

In general, scope differs from invariance in at least two ways. First, invariance is a modal notion: it has to do with whether a relationship *would* remain stable under various hypothetical changes. In contrast, scope, as I understand it, is an "actualist," nonmodal notion: it has to do with how many systems or how many different kinds of systems there actually are for which a generalization holds in the sense described above. Second, insofar as the notion of invariance applies to change-relating generalizations, it requires stability under interventions. Suppose a generalization such as (6.6.3) describes the behavior of springs made of two different kinds of materials,  $M_1$  (plastic) and  $M_2$  (copper). There may be no well-defined notion of changing a spring made of  $M_1$  into one made of  $M_2$  or at least no way of carrying out the change so that (6.6.3) is stable over intermediate steps in the change. (As Hitchcock and Woodward, 2003, put it, (6.6.3) isn't stable in a neighborhood around  $M_1$ .) Hence, one can't legitimately talk about (6.6.3) being invariant under changes from  $M_1$  to  $M_2$ . If (6.6.3) applies to both  $M_1$  and  $M_2$ , it has broader scope than if it applies just to springs made of  $M_1$ , but it isn't any more invariant. As we shall see in chapter 8, the distinction between invariance and scope has important implications for unificationist views of explanation like those defended by M. Friedman (1974) and Kitcher (1989).

## 6.7 Invariance and Exceptionlessness

One of the most important of the traditional criteria for lawfulness is the requirement that laws be exceptionless.<sup>11</sup> Most philosophers still endorse this

idea; a particularly straightforward expression can be found in a recent paper by Paul Pietroski and Georges Rey (1995): “The key feature [of laws] . . . is the universal quantifier. Laws say that whenever some initial condition obtains, some other condition obtains as well. A single instance of  $[F \rightarrow G]$  shows the generalization ‘ $F \rightarrow G$ ’ to be false, in which case a fortiori there is no such law” (p. 83). If taken literally, this requirement virtually forces the law/accident dichotomy on us, because exceptionlessness is an all-or-nothing matter, and not one of degree. This is also the requirement that creates some of the deepest difficulties for the contention that explanatory generalizations in the special sciences are laws because, on the face of things, most such generalizations seem far from exceptionless. The centrality of this requirement in contemporary discussions of laws and explanation is indicated by the existence of a very substantial literature on *ceteris paribus* laws, to be examined in more detail in section 6.14, which is premised on the assumption that to vindicate the lawfulness and explanatory status of generalizations in the special sciences, one must show that these generalizations are (or can be closely associated with “backing” generalizations that are) exceptionless. Needless to say, there would be no motivation at all for this project if (as I argue) it is a mistake to suppose that to qualify as a law or as invariant or explanatory a generalization must be exceptionless.

By way of contrast with the traditional view that exceptionlessness is essential for lawfulness, I said, in describing the ideal gas law (6.2.9)  $PV = nRT$ , that it was invariant under a certain range of changes in the variables  $P$ ,  $V$ , and  $T$ , but broke down or failed to hold exceptionlessly under others (e.g., under conditions of extremely high pressures at which intermolecular forces become important). I thus took it that the ideal gas law could count as a genuine law and figure in explanations of the behavior of gases within the domain over which it is invariant even though it had exceptions outside of this domain. Philosophers who, like Pietroski and Rey, require that genuine laws are exceptionless must hold that the ideal gas law, when formulated as (6.2.9), is no law. The usual move is to suggest that, insofar as there is a law associated with (6.2.9), this will be a generalization that incorporates some appropriate set of qualifications and conditions into its antecedent in such a way as to render it exceptionless. Thus, it will be suggested that the law associated with (6.2.9) is not really expressed by (6.2.9) itself, but by some more complicated, exceptionless generalization such as

(6.7.1) In circumstances  $C_1 \dots C_n$  (where  $C_1 \dots C_n$  are taken to exclude the possibility that intermolecular forces are important, etc.), the ideal gas law (6.2.9) holds.

In fact, however, most known examples of physical laws follow a pattern like the one I have attributed to the ideal gas law. They are invariant only within a certain domain or regime of changes and break down outside of these. For example, the laws of classical electromagnetism (Maxwell’s equations) break down at scales at which quantum mechanical effects become important.



Similarly, the field equations of General Relativity are widely expected to break down at very small distances (the so-called Planck length), at which quantum gravitational effects become important.

In my view, we should resist the conclusion that these facts show that Maxwell's equations and the field equations of General Relativity are not, in their usual formulations, genuine laws and that the genuine laws associated with these generalizations are instead exceptionless generalizations constructed on the model of (6.7.1) (i.e., generalizations like (6.7.2) "If such and such conditions are satisfied, then Maxwell's equations will hold"). To begin with, such a view is sharply at odds with standard scientific practice, which is to take Maxwell's equations and the field equations as laws just as they stand and to appeal to these generalizations, rather than exceptionless reconstructions of them like (6.7.1)–(6.7.2), in order to explain. Moreover, scientists often make use of laws in their usual form (i.e., with exceptions) in circumstances in which they are unable to describe in a precise or theoretically perspicuous way the exact boundaries of the domains over which they are invariant, that is, in circumstances in which they do not know how to turn them into an exceptionless generalizations along the lines of (6.7.1)–(6.7.2). Scientists regarded Maxwell's equations as laws of nature and appealed to them to explain long before they were able to correctly specify the circumstances in which these equations break down. It is a reasonable guess that many generalizations presently regarded as laws similarly will be found to break down in circumstances that are not at present understood. Indeed, some theorists claim that this will be true for all laws of nature—a suggestion that is incoherent on the traditional account of laws, although not on the invariance-based account. Even putting this possibility aside, if we demand that all genuine laws must be exceptionless generalizations, it seems to follow that we know very few laws and, if we make the additional assumption that explanations must cite laws, then most of the generalizations we know how to formulate, even in physics, cannot be used to explain. Finally, and most important, we shall see in section 6.8 that there are principled reasons, grounded in the kinds of information that we expect successful explanations to provide, why generalizations like the ideal gas laws and Maxwell's equations are formulated in their usual, exceptioned form, rather than in the exceptionless form (6.7.1)–(6.7.2).

I believe that it is an important advantage of the notion of invariance that it provides a way of capturing this feature of laws: that it doesn't require that laws be exceptionless. As the ideal gas law (6.2.9) illustrates, it makes perfectly good sense to think of a generalization as invariant across a certain range of changes and interventions even if it is not exceptionless or invariant across *all* changes and interventions. The idea that Maxwell's equations are invariant across certain kinds of changes but not others and that this sort of invariance is sufficient for those equations to count as laws and to figure in explanations represents in a natural way the role that these generalizations actually play in scientific practice.

## 6.8 Independent Specification versus Exception Incorporation

I suggested above that there are principled reasons, in addition to the considerations rehearsed in 6.7, why it is appropriate to think of such generalizations as Maxwell's equations and the field equations of General Relativity as laws just as they stand, and misguided to demand that all laws or explanatory generalizations must be exceptionless. It will be useful to explore these in more detail, as they figure importantly in our subsequent discussion.

Let me begin by raising what might seem to be an obvious objection to the position defended in the previous section. Consider the formulation that I favor, in which we claim that

(6.8.1) Generalization (G) is invariant within a certain domain  $D$ , but breaks down or has exceptions outside this domain.

I call this the *independent specification model* because the domain  $D$  in which (G) holds is specified independently of (G), rather than being packed into the antecedent of (G). The objection to this goes as follows: given the information embodied in (6.8.1), can't we always reformulate this model along the lines suggested in 6.7, as an exceptionless generalization of the traditional kind, the antecedent of which is restricted to  $D$ ? That is, why not think of (6.8.1) as equivalent to an exceptionless generalization  $G^*$  that says

(6.8.2)  $G^*$ : For all  $X$ , whenever  $X$  is in domain  $D$  (or satisfies whatever conditions are sufficient for being in domain  $D$ ), then ... (here follows the original generalization  $G$ ).

I call this the *exception-incorporating model*, because the restrictions on  $D$  are incorporated directly into the generalization (6.8.2) ( $G^*$ ) rather than being specified independently.

Even if it is true that generalizations in science are typically formulated along the lines of the independent specification model (6.8.1), rather than along the lines of the exception-incorporating model (6.8.2), it is tempting to think of (6.8.1) and (6.8.2) as two different ways of representing the same claim, as mere notational variants on one another. Isn't it arbitrary whether we specify the domain independently of (G), as (6.8.1) does, or build it into the antecedent of a more detailed generalization, as (6.8.2) does? If anything, isn't (6.8.2) simpler and more perspicuous than (6.8.1) and more in accord with traditional ideas about the features laws must possess?

In what follows, I argue that the formulations (6.8.1) and (6.8.2) are not equivalent or interchangeable: the two formulations are motivated by quite different views about the information required to improve an explanation and about what one needs to know to successfully explain. They are also associated with two quite different ways of thinking about the content of scientific theories. There are good reasons for preferring (6.8.1) to (6.8.2).

Let me begin with an observation designed to undermine the suggestion that the exception-incorporating formulation (6.8.2) is automatically more natural or perspicuous. Once one gives up the idea that successful explanation is just a matter of nomic subsumption (or of showing that an explanandum was to be expected), it will not be always or automatically true that we improve the explanatory credentials of a generalization with exceptions by replacing it with a generalization that is exceptionless or more nearly exceptionless. In particular, according to the account defended in chapter 5, changes to a generalization that render it more nearly exceptionless but do not enable it to figure in the answers to a larger range of what-if-things-had-been-different questions will not constitute an explanatory improvement. It is in part because of this that exceptionlessness is a less crucial feature of laws and explanatory generalizations than many philosophers have supposed.

As an illustration, suppose that, as we have imagined, General Relativity does break down below the Planck length, but only in those circumstances, so that the generalization

(6.8.3) Above the Planck length . . . (here follow the field equations)

is genuinely exceptionless. Consider some phenomenon such as the deflection of starlight by the sun which we ordinarily take to be explained by General Relativity in its usual, exceptioned form, that is, just by the field equations themselves. Would this explanation be improved if we were to replace the field equations with the exceptionless generalization (6.8.3)? Not (or at least not obviously) according to the account of explanation defended in chapter 5. Taken literally, (6.8.3) tells us nothing about what would happen if the additional condition added to its antecedent—being above the Planck length—were to change. Even if we interpret (6.8.3) as claiming, by implication, that the field equations no longer hold below the Planck length, (6.8.3) certainly fails to provide the kind of precise and specific information about what would happen under this condition that we desire in a successful explanation. Because of this, (6.8.3) tells us little if anything in addition to the field equations themselves about what starlight deflection depends on.

The idea that (6.8.3) does not represent a serious explanatory advance on the field equations as ordinarily formulated may seem very odd to those whose judgments have been nurtured by nomic subsumption models of explanation and by the closely associated idea that laws must be exceptionless. Nonetheless, I submit that this judgment is just scientific common sense. A genuine explanatory advance over General Relativity would require the actual construction of a unified theory of gravity that embraces both quantum and macroscopic gravitational phenomena. Presumably, such a theory would show, in terms of some single set of principles, both how gravitational phenomena behave at very small-length scales and how General Relativity turns out to be correct or nearly correct at large distances. Such a theory could thus be used to answer a larger set of (interesting) what-if-things-had-been-different questions than General Relativity, and for this reason would represent

an explanatory advance. However, merely to specify that GTR breaks down below the Planck length is not to provide such a unified theory (although it no doubt suggests something about the form such a theory will take). It is in part because the substitution of (6.8.3) for the field equations represents no serious explanatory advance that scientists are usually quite happy to employ the usual formulation of GTR rather than (6.8.3) to explain phenomena that fall within the domain of GTR.

To put the point in a way that anticipates my discussion below: the condition “being above the Planck length” plays a different role than the causal or explanatory factors that figure in the field equations such as the stress-energy tensor or the curvature tensor. In contrast to the mass-energy distribution in some region of space-time that genuinely helps to explain why this region has a certain curvature, “being above the Planck length” doesn’t describe a factor that (causally) explains anything or makes anything happen. Its role is rather to describe a condition for the application of GTR or to help specify the domain over which GTR holds. Scientists recognize this different role by not thinking of this condition as built into the field equations themselves, but as specified independently, along the lines of the independent specification model (6.8.1).

There are other considerations, also rooted in scientific practice, that support this analysis. The idea that laws and other explanatory generalizations must be exceptionless goes along with (is supported by) a certain picture of how theorizing and model building work in science: a picture according to which explanatory generalizations already contain (at least if properly formulated) within themselves a full specification of their exact domains of application. However natural this picture may seem to philosophers, it is a descriptively inaccurate account of scientific practice. A more descriptively realistic picture is this: at any given time, scientists will have in their possession various generalizations that they have successfully used to model and explain certain phenomena. However, it is typically a separate empirical question, the answer to which is not already built into these generalizations, what the full range of phenomena is that can be so explained. Sometimes, scientists can give a simple, precise, and general characterization of the domain in which a generalization holds. Then one can read off just from this characterization whether the generalization can be appropriately applied to some potential explanandum. But more commonly, especially in the special sciences, such a general characterization will be unknown and may not even exist at the level at which one is theorizing. Instead, the circumstances in which generalization breaks down will be very complex and heterogeneous, and in many cases not known with any precision. In this sort of case, whether the generalization holds for various previously unexplained phenomena must be discovered empirically, on a case-by-case basis, by seeing whether the generalization can be successfully applied to them, perhaps with the guidance of some rough rules of thumb.

As an illustration, consider Mendel’s law of segregation, formulated as the claim that in sexually reproducing organisms each gene from a pair has 0.5

probability of being represented in a gamete. This generalization breaks down in a number of different circumstances, for example, when meiotic drive is present. However, the law of segregation does not by itself tell us what these circumstances are; instead, whether substantial violations are present in particular populations often must be determined “empirically” on a case-by-case basis. L. C. Dunn (1957), the discoverer that the *t*-allele in house mice (which is responsible for taillessness) does not conform to Mendelian segregation, describes this feature of biological practice when he writes: “Mendelian heredity and its corollary, Hardy–Weinberg equilibrium in panmictic populations, assume [that the probabilities of the *A* and *a* gametes produced by the heterozygote *Aa* are equal] as a matter of course and the assumption is generally justified by direct evidence and by success in application. But the rule is not universal” (pp. 139–40, quoted in Beatty 1979, pp. 131–32). In contrast to the exception-incorporating model, Dunn does not think that the law of segregation is exceptionless or that one can determine whether some particular trait conforms to the law merely by examining whether the conditions specified in the antecedent of the law are satisfied. Instead, such a determination is made on the basis of additional evidence or “success in application.”

As a second illustration, consider the principles of rational choice theory. In addition to the violations of these principles discussed in section 6.4, there is general agreement that exceptions are more extensive in connection with certain kinds of political and economic phenomena than others. For example, in the course of a recent defense of rational choice models, Fiorina (1995) claims that: “RC [rational choice] models are most useful where stakes are high and numbers low, in recognition that it is not rational to go to the trouble to maximize if the consequences are trivial and/or your actions make no difference” (p. 88). In a similar vein, Green and Shapiro (1995) write, in the course of a critical survey of RCT:

Rational choice explanations should be expected, *prima facie*, to perform well to the extent that the following five conditions are met: (i) the stakes are high and the players are self-conscious optimizers; (ii) preferences are well ordered and relatively fixed (which in turn may require actors to be individuals or homogeneous corporate agents); (iii) actors are presented with a clear range of options and little opportunity for strategic innovation; (iv) the strategic complexity of the situation is not overwhelmingly great for the actors, nor are there significant differences in their strategic capacities; and (v) the actors have the capacity to learn from feedback in the environment and adapt. Our conjecture is at bottom empirical, rooted in our best judgment concerning why rational choice models have failed in the literatures we have examined. We might be wrong about one or more of these constraints; only the progress of empirical inquiry will tell. (p. 267)

For example, rational choice models typically provide better explanations and more accurate predictions of the behavior of political elites and party leaders (who are often in a position to exert a strong influence on outcomes about

which they are informed and care a great deal) than of the decisions of individual voters, who are often not well-informed and whose chances of casting a decisive ballot are typically extremely small. For similar reasons, as Satz and Ferejohn (1994) observe in a recent paper, rational choice models have been far more successful in explaining the behavior of firms than the behavior of individual consumers.

In the passages quoted above, the likely domain of RCT and the circumstances in which it is likely to break down are specified in an informal and rather imprecise way and independently of the basic explanatory principles of RCT, rather than being incorporated into those principles. That is, they follow the pattern of the independent specification model (6.8.1) above, rather than the exception-incorporating model (6.8.2). One reason it is implausible to suppose that, appearances to the contrary, these restrictions should be regarded as built into the principles of RCT is that the restrictions are inconsistent with the principles if the latter are interpreted as universal laws. For example, RC principles themselves tell us that we should not expect that people's behavior will be any less self-interested when stakes are low than when they are high. It is precisely because of this that explaining why most people bother to vote creates such difficulties for RC approaches. In addition, the restrictions described above are obviously vague and imprecise; they are best viewed as rules of thumb rather than as specifications of the exact circumstances in which RC principles will hold. This imprecision makes the restrictions unattractive candidates for incorporation into the antecedents of RC principles themselves, a point to which I return below. Finally, as the above quotations make clear, the restrictions represent empirical discoveries that result from a long series of unsuccessful attempts to apply rational choice approaches to the phenomena described above, attempts that clearly reveal the extent to which those who use the theory do not regard a specification of the phenomena for which the theory holds as built into the fundamental principles of the theory.

Although model (6.8.1), in which domain and generalization are specified independently, often seems to provide a more accurate description of scientific practice, one might still wonder whether the exception-incorporating model (6.8.2) is more normatively perspicuous. In what follows, I argue that the reason scientific practice conforms to model (6.8.1) is that this model also has certain normative advantages: it allows us to formulate a much more plausible account of what needs to be known in order to successfully explain. Implicit in the account defended in chapter 5 are the following epistemic requirements: if one wishes to explain the behavior of system  $S$ , one needs to know (6.8.4), an invariant generalization ( $G$ ), and (6.8.5) information about initial conditions holding in  $S$  that, when combined with ( $G$ ), can be used to answer a range of what-if-things-had-been-different questions about the behavior of  $S$ . Knowing this requires knowing (6.8.6) that with respect to the behavior in question,  $S$  is in the domain of invariance of ( $G$ ). However, to use ( $G$ ) to explain, one need not know (6.8.7) the exact boundaries of the domain  $D$  over which ( $G$ ) holds; that is, one doesn't have to know an exceptionless version of ( $G$ ).

This conception fits naturally with the independent specification model (6.8.1) and with the remarks from Florina and Green and Shapiro quoted above. The idea is that one can appeal to the principles of RCT to explain the behavior of, say, buyers and sellers in a certain market, as long as one knows that the behavior of these agents is within the domain of invariance of these principles (i.e., as long as these principles correctly describe how those participants would behave under some relevant range of changes in variables like prices), even though one is unable to state these principles in a completely exceptionless form and they almost certainly break down in unknown ways for some other actors in other circumstances. The fact that such principles are violated by, say, ordinary voting behavior doesn't undercut their use to explain behavior in domains in which they are invariant. Similarly, a nineteenth-century physicist can use Maxwell's equations to explain various classical electromagnetic phenomena while both he and his audience have false beliefs about (or no beliefs at all about) the conditions under which Maxwell's equations fail to hold—while being unable to formulate Maxwell's equations in a genuinely exceptionless way—as long as it is known to be true that these classical phenomena fall within the domain of invariance of Maxwell's theory. The independent specification model permits us to distinguish between, on one hand, knowing (6.8.4)–(6.8.6) (i.e., the putative explanatory generalization  $G$  and that we are within the domain of invariance of  $G$ ), and, on the other hand, knowing (6.8.7) (the exact boundaries of the domain of  $G$ ), because we don't think of the information (6.8.7) as already built into ( $G$ ) and because we can know that (6.8.6) the system of interest is within the domain of ( $G$ ) without knowing the exact boundaries of that domain. It thus allows us to express the idea that it is invariance rather than exceptionlessness that is crucial to successful explanation. By contrast, on the exception-incorporating model, no such distinction is available. Information about the boundaries of  $D$  must be built into ( $G$ ) itself, and if, as is typically the case, one doesn't have such information, one will typically be unable to formulate ( $G$ ) itself in an acceptable way.

Although the exception-incorporating model (6.8.2) connects successful explanation with the possession of exceptionless generalizations, the independent specification model (6.8.1) fits more naturally with the undeniable fact that in the special sciences, we often must appeal to generalizations such that the exact boundaries under which they hold are unknown or difficult to characterize precisely. My suggestion is that it is part of our methodology for constructing and evaluating explanations that this sort of imprecision is allowable in the specification of the domain over which an explanatory generalization holds but not acceptable when specifying the generalization itself. The informal qualitative descriptions above of the domain over which rational choice theory may be expected to hold illustrate this basic idea: the imprecision of such descriptions is acceptable when characterizing the domain of the theory, but would be unacceptable if built into the fundamental generalizations of RCT. When our knowledge of the limits of validity of a generalization are vague, or when we know or suspect it doesn't hold exceptionlessly

but are unable to fully enumerate the exceptions, we build the vagueness into our characterization of its domain rather than building it into the antecedent of the generalization itself. Nor is this arbitrary; as suggested above, we can operate perfectly well with domains with vague boundaries, because often we can know that we are within those boundaries even if we don't know exactly what they are. By contrast, when vague and unclear domain restrictions are built into the antecedent of a generalization, we are left with a candidate for a law without any definite content at all.

## 6.9 Invariance and Counterfactuals

As I have emphasized, the notion of invariance is a modal or counterfactual notion: it has to do with whether a relationship would remain stable if, perhaps contrary to actual fact, certain changes or interventions were to occur. What is the relationship between the idea that laws (and other explanatory generalizations) describe invariant relationships and the more familiar idea, defended by many writers, that what distinguishes laws from accidental generalization is that the former but not the latter "support" counterfactuals? Does the invocation of invariance merely restate this more familiar idea in slightly different language? I have already touched on this issue, but in what follows I explore it in more detail. I argue that the relationship between laws (and other explanatory generalizations) and counterfactuals suggested by the notion of invariance differs in a number of important respects from the standard philosophical picture of their relationship.

According to the traditional view about the relationship between laws and counterfactuals, laws have the form of universally quantified conditional claims such as

(6.9.1) All *As* are *Bs*

and the counterfactuals they support have the following form:

(6.9.2) If this *x* were to be an *A*, then it would be a *B*, where *x* is any arbitrary object, including (especially) those that are not at present *As*.

By contrast, if (6.9.1) is accidental, it will fail to support such counterfactuals. Exactly what "support" means and exactly how the counterfactuals of the form (6.9.2) are to be interpreted is typically left unclear, but in practice, the requirement is often interpreted in such a way that (6.9.1) counts as a law if we can find some true counterfactual (or perhaps some small set of true counterfactuals) to associate with it.<sup>12</sup>

There are several problems with this proposal. First, there seem to be many generalizations that in some sense support<sup>13</sup> counterfactuals of the form (6.9.2), but that are plainly not laws, or even invariant generalizations. Consider a variant of an example due to Aardon Lyon (1977). A museum has adopted a policy such that



(6.9.3) All of the Sisleys in its possession are hung in room 18.

You are ignorant of this policy and ask, regarding some painting in room 17, whether it is a Sisley. You are told in response:

(6.9.4) If this painting were a Sisley, then it would be in room 18.

There is a natural reading of the counterfactual (6.9.4) according to which it is true and according to which it is supported by the generalization (6.9.3). Nonetheless, (6.9.3) is no law and is a dubious candidate for a causal or explanatory generalization.

Indeed, it is easy to imagine circumstances in which even the generalization

(6.9.5) All the coins in Clinton's pocket are dimes

“supports” (or appears to support) some counterfactuals of the form (6.9.2). Suppose (6.9.5) is found to be true over an extended period of time and in circumstances in which many coins move in and out of Clinton's pocket. Then it is a very reasonable inference that there is some systematic reason or cause why (6.9.5) is true. Perhaps Clinton has made it his policy to allow only dimes in his pocket. If so, and you are told that some particular coin *c* is in Clinton's pocket, you have good reason to suppose that it is a dime. You thus have good reason to accept the counterfactual

(6.9.6) If *c* were a coin in Clinton's pocket, then it would be a dime

provided this is understood as a claim about what it would be reasonable to believe about *c*, given the information that it is in Clinton's pocket, rather than as a claim about what would happen to a nondime if it were introduced into Clinton's pocket as a result of a testing intervention.

These examples illustrate one important difference between the invariance-based account and the standard view of the relationship between laws and counterfactuals: in contrast to the standard view, the invariance-based view attaches a special significance to a particular sort of counterfactual, counterfactuals that describe what would happen to *B* under *interventions* that bring about *A*. As the above examples illustrate, a generalization can support some counterfactuals without supporting any counterfactuals describing what will happen under interventions. For example, although (6.9.3) supports some counterfactuals, it would break down under a testing intervention that consists in introducing a Sisley into room 17 or under other changes in the museum's policy regarding the hanging of pictures. More generally, (6.9.3) fails to support any counterfactuals like

(6.9.7) If one were to introduce a Sisley into the museum via a testing intervention, then it would be in room 18.

Parallel observations hold for (6.9.4). As already argued, it will break down under testing interventions that consist of the introduction of a nondime coin into Clinton's pocket.

Broadly speaking, the examples just described seem to work in the following way. We have a generalization that holds because certain background conditions or generating structures are in place (e.g., a decision to hang all Sisleys in room 18 or the policy on Clinton's part to allow only dimes in his pockets). If we have good reasons for thinking these background conditions will persist, we also have good reasons to think these generalizations will continue to hold and hence good reasons to accept counterfactuals like (6.9.4) and (6.9.6) when they are interpreted in a noninterventionist way.<sup>14</sup>

The requirement that invariant generalizations support counterfactuals about what would happen under testing interventions is closely connected to another important difference between the standard view of the relationship between laws and counterfactuals and the invariance-based approach (cf. Hitchcock and Woodward, 2003). Suppose that we wish to explain the behavior of some object or system  $o$ . As the standard view is usually understood, it claims that generalizations of form "All  $A$ s are  $B$ s" are explanatory of the behavior of  $o$  if they support counterfactuals of the following form:

(6.9.8) If some object  $o^*$ , different from  $o$  and that does not possess property  $A$ , were to be an  $A$ , then it would be a  $B$ .

Call such counterfactuals "other object" counterfactuals: they describe what the behavior of objects other than  $o$  would be under the counterfactual circumstances in which they are  $A$ . By contrast, according to the view I have been defending, to count as invariant and hence explanatory with respect to  $o$ , a generalization must support "same object" counterfactuals that describe how the *very object*  $o$  would behave under an intervention. In particular, the counterfactuals that must be supported by a generalization that is explanatory with respect to  $o$  are of the following form:

(6.9.9) If the value assigned by the variable  $X$  to  $o$  were to be changed via an intervention (e.g., from  $X(o) = 0$  to  $X(o) = 1$ ), then the value assigned by  $Y$  to  $o$  would change in some way predicted by the generalization.

I maintain that it is the ability of a generalization to support same-object counterfactuals concerning interventions (counterfactuals of the form (6.9.9)), rather than its ability to support other-object counterfactuals (counterfactuals of form (6.9.8)), that determines its explanatory status. As an illustration, consider (5.1.3) in which Coulomb's law is used to explain why a conductor with a certain geometrical configuration and charge distribution produces an electric field of strength  $E$ . According to the traditional view of the relationship between laws and counterfactuals, Coulomb's law is considered a genuine law (and hence explanatory) because it supports counterfactuals about how other objects, including nonconductors, would behave if they were to become conductors. Taken literally, this would require us to think of Coulomb's law as supporting such counterfactuals as "If Bill Clinton, the government of Italy, or a neutron were a conductor with such and such a geometrical configuration and charge distribution, he/she/it would produce an electric field of strength  $E$ ." Such counterfactuals are, to say the least, hard to understand and evaluate.

Moreover, they tell us very little about what the field depends on, both because they are difficult to comprehend and because, to the extent we can understand them, they involve changes in the *identity* of the conductor, and this is *not* a factor on which the strength of the field depends. Instead, the strength of the field produced by a conductor depends on such factors as its geometry and charge density.

By way of contrast, my view is that Coulomb's law can be used to explain the field produced by some particular conductor because it supports counterfactuals about what would happen if the charge density or geometric configuration of that very conductor were changed in various ways as a result of interventions. Unlike the other-object counterfactuals considered above, these same-object counterfactuals make explicit how the strength of the potential field produced by the conductor depends on its geometry and charge distribution. To the extent that explaining an outcome or phenomenon is a matter of explaining what it depends on, it is same-object counterfactuals involving interventions rather than other-object counterfactuals that are relevant. Same-object counterfactuals involving interventions are typically clear enough in meaning, and we often have or can obtain scientific evidence that is relevant to their truth. By contrast, many other-object counterfactuals lack a clear interpretation in terms of interventions and seem odd or incoherent or lack any scientific basis.

The focus of the traditional view on other-object counterfactuals is, of course, closely connected to the traditional idea that laws or explanatory generalizations must be exceptionless: if laws or explanatory generalizations are required to support other-object counterfactuals, they must specify nontrivial conditions that ensure that any arbitrary object meeting those conditions will behave in a certain way (e.g., conditions that are sufficient to ensure that any arbitrary object or system conforms to Coulomb's law). By contrast, if explanatory generalizations are required only to support same-object counterfactuals, they may correctly describe what would happen under interventions in some system of interest while failing to correctly describe (or failing to apply to) the behavior of other systems (e.g., quantum mechanical systems for which Coulomb's law and other laws of classical electromagnetism fail). The idea that same-object counterfactuals are what matter from the point of view of explanation thus goes along with the idea (which I mean to endorse) that the explanatory status of Coulomb's law with respect to macroscopic systems like the one described in (5.1.3) is not impugned by the existence of other systems (e.g., quantum mechanical systems) for which that law fails to hold, as long as it is the case (as it clearly is) that the law correctly describes (up to some reasonable degree of approximation) how the electromagnetic field will change under interventions in the sorts of systems whose behavior we are seeking to explain.

It is ironic that the traditional conception of laws is often defended on empiricist grounds: exceptionless generalizations that support other-object counterfactuals are thought to be empirically respectable, whereas generalizations having a richer modal content such as those making claims about what would happen under interventions are not. In my view, this is exactly backwards.

The interventionist, same-object counterfactuals that are central to the manipulationist account I have been defending are typically clear enough in meaning, and we often can obtain scientific evidence that is relevant to their truth. For example, it may be possible to experimentally intervene to change the geometry or charge distribution of the conductor to determine whether the relationship expressed by Coulomb's law continues to hold for the conductor. On the other hand, many other-object counterfactuals lack a clear sense or any empirical basis. It is wholly mysterious how we might test counterfactuals about what would happen if Bill Clinton, the H.M.S. *Victory*, or a neutron were to be a long, straight wire. The difference between these two sorts of counterfactuals is reflected in the judgments of most scientists about the physical content of Coulomb's law. It is natural and intuitive to think of that law as telling us how the field due to some conductor would change as its charge density or geometry were changed, but the law is simply not in the business of trying to tell us what would happen under the fantastic transformations contemplated above.

As another illustration, assume for the sake of argument that

(6.9.10) All dry, well-made matches ignite when struck

is a law or at least a true causal generalization. The oddity of the view that this generalization should be understood as supporting counterfactuals about what would happen if any arbitrary object were to satisfy its antecedent is very memorably brought out in a wonderfully titled paper by Adam Morton (1973), which asks, in effect, whether (6.9.10) supports the counterfactual

(6.9.11) If I were a dry, well-made match, I would ignite if struck.

Opinions may differ about the truth value or meaningfulness of (6.9.11), but it seems uncontroversial (this is the main point of Morton's paper) that counterfactuals like (6.9.11) are the wrong counterfactuals to look at in evaluating whether (6.9.10) is a true causal or (in the relevant sense) a "counterfactual supporting" generalization. Rather than having to do with how objects that are not matches would behave if they were to become matches, the modal or counterfactual import of (6.9.10) instead has to do with the contrast in behavior with respect to ignition between those dry matches that are struck and those matches that are not. There is arguably no such thing as a coherent, well-defined intervention that would transform a human being into a match, but there are well-defined interventions that involve striking a match and sensible counterfactuals about what would happen under such interventions that are testable in principle and often in practice. It is such interventionist counterfactuals about what would happen to some particular match when struck that are relevant to explaining its behavior.

I turn now to a second difference between the traditional view of the relationship between laws and counterfactuals and the account I advocate. On the traditional view, one asks a question that requires a single yes or no answer. Either one considers just a single counterfactual of the form

(6.9.2) If  $X$  were to be an  $A$ , then it would be a  $B$

and asks whether such a counterfactual is supported by (6.9.1) (All *As* are *Bs*), or else (less commonly) one considers all counterfactuals of the form (6.9.2) and asks whether all are supported by (6.9.1). In either case, a generalization either passes this test and hence qualifies as a law or else it fails the test and is accidental. (This reflects the dichotomous character of the traditional law/accident framework.) By contrast, as I have emphasized, a generalization can be more or less invariant and it can be invariant under one set of interventions or changes and not under others. Instead of associating a single test for counterfactual support with a generalization like (6.9.1), the invariance-based approach suggests that we think in terms of a whole family of counterfactuals whose antecedents correspond to different interventions or different changes in background circumstances under which *A* would be true and then ask whether *B* would also be true in those circumstances. Some of these counterfactuals may turn out to be true and others false, and together they give us the range of circumstances or interventions over which the generalization is invariant. We thus do not confine our attention to a single world or set of worlds that is “closest” (however this is defined) to the actual world in which *A* is true and ask which *B* is true in that world, but rather regard it as legitimate to consider counterfactuals in which *A* occurs under conditions that are very dissimilar from those that hold in the actual world. For example, in considering the range of changes over which the field equations of General Relativity are invariant, it will be appropriate, on an invariance-based approach, to consider worlds in which the mass distribution of the universe is extremely different from that obtaining in the actual world (e.g., worlds consisting of a single star or that are entirely empty of mass) and to ask whether the equations of GTR would continue to hold under such circumstances. This is reflected in the standard practice of relativists, which often assumes that those equations are highly invariant in the sense that they would continue to hold under such extreme circumstances.

Similarly, consider the automotive generalization (6.4.1) relating gas pedal position and speed. As we have noted, this generalization supports some counterfactuals concerning what would happen under interventions that change the position of the pedal in circumstances that are close to the actual circumstances (similar environmental conditions, similar internal state of the engine, etc.). However, in assessing the range of invariance of (6.4.1), we also consider what would happen under cases in which the internal structure of the car and other features of its environment are changed in ways that depart much more radically from the actual circumstances. It is the fact that we think that (6.4.1) would break down under many such counterfactual circumstances that distinguishes it from less fragile generalizations such as the field equations. More generally, rather than focusing on a single counterfactual of the form (6.9.2), we focus on a whole family of counterfactuals corresponding to different ways of strengthening the antecedent of (6.9.2): we consider a whole set of counterfactuals of the form

(6.9.2\*) If *A* and *C* were true, then *B* would still be true

where different possible interventions bring about  $A$  and  $C$  represents different possible background circumstances in which  $A$  might occur.

## 6.10 Invariant Generalizations That Are Not Laws

I suggested above that there are invariant generalizations that are not naturally regarded as laws, if one has even a modestly demanding conception of what a law is. As an illustration, consider again the example from section 6.6 of a particular sort of spring  $S$  that, over a certain range of extensions, conforms to Hooke's law (6.6.3)  $F = -K_s X$ , where  $X$  is the extension of  $S$ ,  $F$  the restoring force it exerts, and  $K_s$  a constant characterizing  $S$ . Suppose that (6.6.3) is invariant under some interventions that change the extension of  $S$ , but breaks down for extensions that are too large and for other sorts of changes in background conditions as well. As we have seen, this fact by itself does not distinguish (6.6.3) from paradigmatic laws of nature such as Maxwell's equations. Nonetheless, there appear to be several respects in which (6.6.3) does differ from paradigmatic laws. First, as already discussed, (6.6.3) is much narrower in scope. A second difference, which is perhaps more significant from the point of view of explanation, is this: not only will there be a range of "extreme" extensions for which (6.6.3) breaks down, but even if we confine our attention to extensions of  $S$  that are not in this range—that is, extensions for which (6.6.3) sometimes holds—there will be a large number of possible changes in background conditions that do not explicitly figure in (6.6.3) that will result in a violation of (6.6.3). For example, even if it is true that in "normal" circumstances  $S$  will conform invariantly to (6.6.3) under small extensions, it will not do so if it is heated to a high temperature or cut with shears. Similarly, (6.6.3) will break down if we intervene to produce even a small extension in the wrong way; for example, if the intervention physically deforms the spring. It matters how we produce a given value of  $X$  in (6.6.3) and not just what that value is. In general, the set of possible "interfering conditions" for (6.6.3) is very large and heterogeneous and will resist any simple, informative characterization. Because we don't know how to characterize all the items in this set in a noncircular, illuminating way, we find ourselves saying things like the following: (6.6.3) holds if "other things are equal" or in the absence of "disturbing factors," where no very precise independent specification of the quoted phrases is available. Similar observations hold for a number of other generalizations considered above, for example, for the generalization (6.4.1) linking position of the gas pedal and maximum speed for a particular kind of car.

By way of contrast, although paradigmatic laws like Maxwell's equations do break down under certain extreme values of the variables figuring in those equations, whether the equations hold or not depends just on the values of those variables and not on how those values are brought about. That is, in contrast to the Hooke's law generalization (6.6.3), within the classical regime for which Maxwell's equations hold, it does not matter how we change the

distances between point charges, or the intensities of electromagnetic fields, and so on: Maxwell's equations will continue to hold under such changes. Moreover, changes in background conditions play a different role in connection with the invariance characteristics of paradigmatic laws than in connection with generalizations like (6.6.3). When the circumstances under which paradigmatic laws fail to be invariant are known, they typically can be given a relatively simple, unified characterization. Such circumstances seem to fall into one of two categories: laws break down either for extreme values of variables that explicitly figure in them (e.g., high temperatures and pressures, in the case of the ideal gas law) or when some very small set of variables that have been omitted from the law diverge from a limiting value—the pattern being that the law holds when the variables take this limiting value but not otherwise. For example, according to a well-known textbook on General Relativity, the Newtonian inverse square law “is an excellent approximation in the limiting case of low velocity in a weak gravitational field” (Ohanian 1976, p. 2). That is, the law breaks down both when gravitational fields are strong (an extreme value of an included variable) and also when an omitted variable (velocity) is not small in comparison with the speed of light.

On this way of looking at matters, the differences between the Hooke's law generalization (6.6.3), on the one hand, and paradigmatic laws like Maxwell's equations on the other, although real, look very much like differences in degree (of scope and of range of interventions and changes in background conditions over which these generalizations are invariant) rather than of kind. Paradigmatic laws are simply generalizations with wide scope that are invariant under a large and important set of changes that can be given a theoretically perspicuous characterization. We are willing to regard other invariant generalizations as laws to the extent that they resemble these paradigms in these respects. It is thus not surprising that the boundary between those invariant generalizations we regard as laws and those we do not is fuzzy and contentious—an additional reason for resisting models of explanation or accounts of causation that require a sharp law/nonlaw boundary.

These considerations in turn raise an obvious question: Given that the difference between Maxwell's equations and (6.6.3) is one of degree, why not reflect this continuity by extending the notion of “law” to cover all generalizations that are invariant under some interventions and changes in background conditions, so that generalizations like (6.6.3) and the generalization (6.4.1) linking pedal depression and automobile speed count as laws as well, albeit local or qualified or *ceteris paribus* laws? In effect, many writers have proposed that we do just this (Hausman 1992; Fodor 1991; Kincaid 1989).

In thinking about this proposal, it is important to separate issues that are largely terminological, in the sense that they reflect decisions about how to use the word “law,” from more substantive issues. To the extent that the proposal under discussion accepts my claims about the importance of invariance and its role in explanation and simply extends the word “law” to cover all invariant generalizations, it differs only verbally from my own position. In fact, however,

few if any philosophers who want to extend the notion of law have in mind only such a terminological proposal. Instead, philosophers who have thought of generalizations like the Hooke's law generalization (6.6.3) (or the various generalizations of the special sciences) as laws have usually been motivated by a very different account from mine of the features that make such generalizations explanatory. They have tried to show that such generalizations are explanatory in virtue of satisfying (or to the extent they satisfy) various of the traditional criteria for lawfulness rather than in virtue of being invariant or figuring in the answers to a range of what-if-things-had-been-different questions. For example, many treatments of so-called *ceteris paribus* laws, discussed in section 6.13 below, are motivated by the idea that if a generalization is to be a law and hence explanatory it must itself be or be "backed" by an exceptionless generalization. The philosophers who advocate such treatments are not merely proposing that we extend the word "law" to cover the explanatory generalizations of the special sciences but are instead adopting a distinctive substantive position about the features (namely, exceptionlessness) that such generalizations must possess if they are to figure in explanations.

It is in part for this reason—that in practice, the project of extending the notion of law to cover the generalizations of the special sciences is closely bound up with the misguided project of trying to show that, despite appearances to the contrary, these generalizations are explanatory because they satisfy (at least many of) the traditional criteria for lawfulness—that I think clarity is best served by adopting a more restricted notion of law. Moreover, there are additional reasons for such a restricted notion. First, although there are, as I have argued, important continuities between generalizations like Maxwell's and the Hooke's law generalization (6.6.3), there are also very real differences. These may be matters of degree, but they are not for that reason unimportant. The features possessed by generalizations like Maxwell's equations—greater scope and invariance under larger, more clearly defined, and important classes of interventions and changes—represent just the sort of generality and unconditionality standardly associated with laws of nature. Their relative absence from generalizations like (6.6.3) and from many of the explanatory generalizations of the special sciences makes it misleading to assimilate these to paradigmatic laws.

Second, and more important, if the argument of this essay about invariance and explanation is correct, there is no real motivation for such an assimilation. The claim that the explanatory generalizations of the special sciences are laws would have an obvious motivation if there was some independent reason for supposing that all explanation requires laws, understood along the traditional lines. It would also have an obvious motivation if there was some independent reason to suppose that all generalizations must fall into one of two mutually exclusive categories: the lawful or the purely accidental. However, I have argued that we should reject these assumptions. Both are gratuitous once we accept the alternative account of explanation described in chapter 5. Once we accept this alternative account, we don't need to argue that the generalizations of the special sciences are laws (thereby incurring the burden of claiming that



they do not differ in important respects from Maxwell's equations and that, appearances to the contrary, they satisfy such traditional criteria as exceptionlessness) to vindicate their explanatory status. Finally, as will become clear below (see especially section 7.8), the more restricted usage that I favor also has the advantage that it captures what seems to be at stake when philosophers and scientists deny, as they frequently do, that the explanatory generalizations of the special sciences are laws. Writers who take this position typically are not merely making a proposal about terminology; instead, they think that there are important differences between generalizations like Maxwell's equations and many of the explanatory generalizations of the special sciences, differences that a descriptively adequate account of explanatory practice in different areas of science should aim to capture. The framework I have proposed allows us to do this.

### 6.11 The MRL Theory

I turn now to what is perhaps the most widely accepted account of laws: the so-called Mill–Ramsey–Lewis (MRL) theory. A fully satisfying discussion of the MRL theory requires much more detail than I am able to provide here; my remarks focus mainly on some of the differences between this theory and the invariance-based account that I advocate.

A widely discussed formulation of the MRL theory is provided by David Lewis ([1973] 1986b): "A contingent generalization is a *law of nature* if and only if it appears a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength" (p. 73). Here, strength is a measure of informativeness: a theory is stronger the "more" consequences can be deduced from it.

One possible line of criticism of this idea focuses on the role it assigns to simplicity: the very general, domain-independent notion of simplicity required by the MRL theory has never been made clear, it is far from obvious that anything resembling this notion plays the sort of role in theory choice assigned to it by the MRL theory,<sup>15</sup> and it is also dubious that there is any objective, nonarbitrary basis for choosing a best combination or optimal trade-off of simplicity and strength. However, rather than pursuing these criticisms, I want to focus instead on the role of strength. In some cases, this notion seems straightforward: if the set of deductive consequences of  $T_1$  is a proper subset of the set of deductive consequences of  $T_2$ , then obviously  $T_2$  is stronger than  $T_1$ . However, this consideration generates only a partial ordering. The question that then arises is how to understand the notion of strength in other sorts of cases.

Consider a world  $W_1$  in which some regularity  $R$  is instantiated just once. Now compare this with world  $W_2$  in which  $R$  is instantiated many times. On one very natural understanding of strength, a system of axioms that does not allow one to deduce  $R$  in world  $W_1$  sacrifices relatively little in the way of strength. There is only one particular fact—the single instantiation of  $R$ —that

one will be unable to deduce. If the only way to provide for the deduction of  $R$  was to add some relatively complicated axiom  $A$  that permitted the derivation of no other regularities, we might reasonably judge that the gain in strength from the addition of  $A$  would be outweighed by a loss in simplicity. But in  $W_2$  it looks as though the addition of  $A$  will allow one to deduce much more about what actually happens: the occurrence of each of the many instances of  $R$ . If, as seems plausible, this represents a much larger gain in strength in connection with  $W_2$  than in connection with  $W_1$ , there will be a much stronger case that in  $W_2$  the gain in strength that would result from adding  $A$  will outweigh the loss in simplicity. With enough instances of  $R$ , produced perhaps by a very diligent experimenter, we may very well reach a point at which  $A$  qualifies as an axiom in the best deductive system. But it seems, to put it mildly, paradoxical that I (or nature) can strengthen the case that  $A$  (or  $R$ ) is a law by piling up instances in this way. This example points up how different the conception of lawfulness captured by the MRL account is from the notion of invariance, for, as we have seen, the invariance of a generalization has nothing to do with the number of instances it has; number of instances is instead related to what I called scope in section 6.6.

On the other hand, it may be that the notion of strength should be understood differently, in such a way that what matters to strength is not the number of instances of a regularity but the number of "different" regularities that can be deduced. On this interpretation, any generalization having at least one instance counts as a description of a regularity, and if we can deduce such a generalization from more fundamental assumptions, this will count just as much as a gain in strength as the deduction of a regularity with many instances.

This proposal again leads to a very different account of laws than the invariance-based account. As argued in 6.6, not only does the invariance of a generalization have nothing to do with the number of instances it has, it also has nothing to do with the number of different kinds of instances it has or the number of different, more specific regularities it subsumes. There are also a number of difficulties with this new proposal. Once the notion of a regularity is severed in this way from the idea of a repeated pattern in nature with a number of instances, the notion becomes quite unclear. (Notoriously, any claim about a particular matter of fact, such as "Socrates is white," may be treated as a generalization that describes a regularity under this permissive notion of regularity; see Hempel 1965b, pp. 264ff, 340ff). In addition to this, if the suggestion is to yield more than a partial ordering, we require some way of counting regularities, some way of deciding whether we have one regularity or several different ones. It seems dubious that there is any principled way of doing this.

As an illustration of this last point, consider Kepler's (three) laws and Galileo's law of freely falling bodies. It might seem natural to regard these as four distinct regularities; indeed, it is usually taken to be a very impressive achievement of Newtonian mechanics and gravitational theory that they permit the unified explanation of these apparently very different generalizations which were previously thought to be unrelated. On the other hand, the

very fact that these regularities are so explainable immediately raises the question of whether, from the point of view of fundamental science, they are really distinct regularities. Perhaps instead we should see them all as instances of the same fundamental regularities—namely, those of Newtonian mechanics and gravitational theory—working on different initial conditions. Indeed, one suspects that if the history of science had been different and if Kepler's laws and Galileo's law had not been formulated prior to Newton's discovery, this point of view—that they are really the same regularity—might strike us as very natural. It is far from obvious that there is anything in physics itself that tells us that we should regard these as four different regularities. It seems particularly unlikely that the judgment that these are four regularities can be based on some general non-domain-specific principle for counting number of regularities. However, it is just such a principle that seems to be required if number of regularities is to be a useful measure of strength within the MRL framework.

Thus, however we understand the notion of the strength contributed by the addition of a regularity to a deductive systemization—whether as a function of the number of instances that it has or as a function of the number of different kinds of instances it has (or as a function of the number of different kinds of more specific regularities subsumable under it)—it seems to have counterintuitive consequences.

I turn now to another (related) feature of the notion of strength that seems problematic. Consider two worlds,  $W_1$  and  $W_2$ . In  $W_1$ , given the regularities,  $S_1$  is the best deductive system.  $W_2$  is just like  $W_1$  with respect to the regularities that actually occur in it, until time  $t$ . At  $t$ , a scientist proposes a theory  $S_2$  that differs from  $S_1$  in what it predicts what would happen if certain experiments were done or possible interventions were carried out. (We may suppose that  $S_1$  either makes no predictions at all about what would happen under these eventualities, or else makes definite predictions but that these differ from those of  $S_2$ .)

Now consider two possibilities. Under possibility (i), these experiments are not actually carried out. As a result, the regularities in  $W_2$  continue to be just those that hold in  $W_1$ , and the laws of  $W_2$  are the axioms and theorems of  $S_1$ . Under possibility (ii), the experiments are carried out and result in new regularities that are just those predicted by  $S_2$ . Suppose, in fact, that given these additional regularities,  $S_2$  is now the deductive system that achieves the best combination of simplicity and strength with respect to the regularities in  $W_2$ . We may even suppose that this is why our scientist proposed  $S_2$ : she supposed, correctly, that if the experiments were carried out,  $S_2$  would be the “best” deductive system. We thus arrive at the counterintuitive result that, on the MRL theory, what the laws are in  $W_2$  seems to depend on whether the experiments are done. Or, to put what is the same point in a slightly different way, we seem forced to the conclusion that if the experiments are carried out so that the regularities in  $W_2$  differ from those in  $W_1$ , then  $W_2$  and  $W_1$  must necessarily be governed by different laws. But it seems perfectly possible that  $W_1$  and  $W_2$  are governed by exactly the same laws; it is just that in  $W_1$  certain experiments that would allow us to tell whether the laws are those

associated with  $S_1$  or with  $S_2$  have not been carried out, whereas in  $W_2$  they have been carried out. Presumably, this is how our hypothetical scientist will view the matter: she will suppose that there is a preexisting nomological structure common to  $W_2$  and  $W_1$  and will regard her experiments as a way of finding out whether this structure is captured by  $S_1$  or  $S_2$ . She will not suppose that the laws in  $W_1$  are different from those in  $W_2$  merely because experiments are carried out in  $W_2$  that are not carried out in  $W_1$ .

It is perfectly true that what it is *reasonable to believe* about the laws in  $W_2$  or what we can claim to know about them may depend on whether the experiments are carried out. If the experiments are not carried out, it may seem gratuitous to believe that the laws governing  $W_2$  are those of  $S_2$  rather than those of  $S_1$ . But the MRL theory is not supposed to be a theory about what regularities it is reasonable to believe are laws in  $W_2$ ; rather, it is supposed to be a theory about what the laws *are* in  $W_2$ . And when the MRL theory is understood as a theory of what laws are, the result just described seems to show that there is something fundamentally misguided about the whole approach.

One might put the difficulty that we have just exposed in a more abstract way as follows. The MRL account ties the question of what the laws are far too closely to what actually happens or, to be more precise, to which Humean regularities are realized and to how frequently they are realized. This information is certainly part of our *evidence* for what the laws are, but according to our usual conception of law, this information may fail to reveal the full set of nomological relationships holding in some world, roughly because the initial conditions in that world are not such that they allow for the realization (in the form of regularities) of certain lawful possibilities. Thus, in the example immediately above, as we ordinarily think about the matter, we suppose that the laws of  $W_1$  may be those described by  $S_2$  but that the regularities of  $W_1$  may fail to reveal that this is the case because certain initial conditions—those that would obtain when the experiments in  $W_2$  are carried out—fail to obtain in  $W_1$ .

By contrast, whether a relationship is invariant is much less closely tied to what actually happens. If the laws of  $W_1$  are invariant relationships, then we should expect that those laws would continue to hold as various changes in initial conditions and interventions occur. This means, in effect, that these laws will also hold in worlds characterized by different initial conditions besides those that obtain in  $W_1$ —for example, those that obtain in  $W_2$ . At the very least, on the invariance-based approach, we should not conclude that the laws in  $W_2$  are different merely because the initial conditions and hence the regularities are different.

The difference between the sorts of considerations that are relevant to determining whether a generalization is invariant and whether it is a law according to the MRL account can also be brought out by means of some other examples. Consider a world in which a generalization  $G$  has wide scope and is exceptionless. If  $G$  is simple and if one can use it, either alone or in conjunction with the (other) generalizations that qualify as axioms or theorems in the deductive system that best combines simplicity and strength, to deduce lots of

other true generalizations, then it is a good candidate for a law within the MRL framework. However, the fact that  $G$  possesses these features doesn't seem, on the face of things, to have much to do with whether it is invariant. It looks as though a generalization might possess the features just described and yet not be invariant at all or relatively noninvariant. It might depend very sensitively on initial or background conditions or break down under many or all interventions.

In fact, an example described by van Fraassen (1989, pp. 46–47) has just this sort of structure: he imagines a world in which the regularities are those of Newtonian mechanics and Newtonian gravitational theory and that consists of golden spheres moving in stable orbits around each other, iron cubes on the surface of these spheres, and nothing else. Van Fraassen suggests, plausibly, that within the MRL framework,

(6.11.1) All golden objects are spherical

will qualify as a law. This generalization is certainly simple, and without it we would be unable to deduce a major uniformity. Yet the facts about (6.11.1) just described do not at all show that it is invariant; as van Fraassen remarks, nothing about the above story suggests any reason to think that in this world, changes that would have produced nonspherical golden objects are impossible.

Although this example is an artificial one, it is easy to produce realistic illustrations of the same point. Consider again the cosmological uniformities described in section 6.4 which have to do with the large-scale homogeneity and isotropy of the universe. Assume, for the sake of argument, that these regularities hold in our world. The generalizations describing them seem simple, much simpler than generalizations describing a heterogeneous and nonisotropic universe. Moreover, when conjoined with other laws, they can be used to derive a great deal of what we observe; it was on exactly these grounds that they recommended themselves to cosmologists in the first place. On the basis of intuitive judgments of simplicity and strength, we thus have a plausible case that these cosmological generalizations qualify as laws within the MRL framework. Yet, as we have already observed, these generalizations are noninvariant: they would be disrupted if various initial and background conditions were changed. For just this reason they are not regarded as laws by most cosmologists. The moral seems to be the same as in van Fraassen's example: the generalizations picked out as laws by the MRL criteria need not be invariant. Balancing simplicity and strength just doesn't get us invariance.

There is another consideration that reinforces this conclusion. The best deductive systemization criterion is global in character. How, if at all, the presence of some particular generalization  $G$  as an axiom contributes to the overall simplicity and strength of some deductive system  $D$  depends on what other axioms are present in  $D$  (and on how these can be combined with  $G$  to deduce facts about regularities). One cannot judge this just by looking at  $G$  in isolation. By contrast, whether or not a generalization is invariant depends on more local considerations. This is perhaps clearest in cases in which the invariance or lawfulness of a generalization can be established experimentally, although I believe that a similar observation holds for laws or invariant

generalizations the evidence for which is not or is only partially or indirectly experimental, as is presumably the case with many fundamental laws. As an example in which much of the evidence for the invariance of a generalization is experimental, consider the Boyle-Charles law. As explained in chapters 2 and 3, experimentation requires background information of various sorts, but the required information is relatively local and self-contained. We do not think (and experimental practice does not reflect the assumption) that in assessing whether the Boyle-Charles generalization is invariant we need information about how that generalization fits with all the other regularities in the universe. Nor is it very plausible that what investigators *mean* when they claim that the Boyle-Charles law is a genuine law is captured by the MRL account; instead, the invariance-based account does a much better job of capturing the content of this claim.

Although the Boyle-Charles law is an invariant generalization, it is at best a derivative, rather than a fundamental law. Nonetheless the epistemological puzzle it poses also arises for more fundamental laws within the MRL account. Even if direct experimental considerations play a less central role in establishing many fundamental laws, it is nonetheless true that the evidence and background knowledge that do play this role look far more local than one would expect from the MRL account. Consider the question of how a seventeenth- or nineteenth-century investigator might have established that the Newtonian inverse square law or Maxwell's equations were genuine laws on the basis of then available evidence or background knowledge. Let us assume, for the sake of argument, that this was (or might have been) accomplished by looking at information about known regularities and available information about other conjectured laws and judging that as far as this information went, Newton's and Maxwell's generalizations were axioms of an overall deductive systemization that achieved the best combination of simplicity and strength. Even if we grant this assumption, which seems rather dubious, there is a further difficulty. Our imagined researcher has access only to a very limited number of regularities out of the much larger set of "all" regularities in the world and will be unaware of many of the generalizations that will later be regarded as laws. If she applies the MRL account directly to the very limited body of information available to her, it would seem, on the face of things, that she is likely to arrive at a set of candidates for laws that are very different from those a twenty-first-century investigator would arrive at by applying the MRL account to the very different information available to her and that these in turn will be quite different from the candidates an omniscient being with unlimited deductive powers would arrive at by applying the MRL account to "all" the regularities in the world taken at once. What looks like a good candidate for an axiom or theorem in a deductive system that achieves a best combination of simplicity and strength when applied to some small corner of the universe might look like nothing of the kind when the MRL framework is applied more globally.

To put the point in a slightly different way, if what makes the Boyle-Charles generalization or the Newtonian inverse square law a law has to do with the way it fits together with a large number of other generalizations, many of them

describing very different regularities that were unknown at the time that the Boyle–Charles law or the inverse square law was discovered, how is it that procedures involving relatively local kinds of evidence and background knowledge that appear to take little or no notice of these other regularities nonetheless were successful in identifying these generalizations as laws? On what basis, if any, were Boyle or Newton or Maxwell entitled to believe, given the rather local information available to them and knowing very little about the fundamental regularities governing the world, that they had successfully identified generalizations that are axioms or theorems of the systemization (of *all* of the regularities in the universe) that achieves a best combination of simplicity and strength? And why do we not see, when we look at the history of science, a great deal of discontinuity in our judgments of which generalizations are laws, as new information about regularities and new possible systemizations become available to us?

John Earman (1993), a prominent defender of the MRL account who is sensitive to these epistemological worries, puts them as follows:

On the MRL account, whether an individual statement  $L$  expresses a law cannot be determined by features of  $L$  itself. So, for example that  $\nabla \times \mathbf{E} = -\partial \mathbf{B} / \partial t$  expresses a law of electromagnetism depends on the fact that it fits harmoniously together with three other differential equations to form what are called Maxwell's laws. The objection is that this observation backfires. For on the MRL account scientists shouldn't have accorded Maxwell's equations the honorific of "law"; they should have waited to see how these equations fit together with other equations to form a comprehensive system for all of physics. (p. 417)

He replies as follows on behalf of the MRL account:

The response starts from the obvious: given the kinds of creatures we are and given the complexities we face, we can't investigate everything at once but have to focus on selected aspects of the world. We hope that we have managed to focus on a domain of phenomena that is fundamental in the sense that the "laws" we construct on the basis of ignoring everything outside the domain will survive in some recognizable form as laws as we extend the scope of investigation. The history of science for the last hundred years shows that "Maxwell's laws" are robust in this regard. (p. 417)

Earman claims, I think correctly, that from the perspective of the MRL account, the conclusion of nineteenth-century physicists that Maxwell's generalizations were indeed laws was based on the "hope" that future changes in our best total deductive system would preserve the status of these generalizations as axioms or theorems. Earman notes that, given the assumption that the MRL account is correct, this hope has been vindicated so far. But the issue, as I see it, is not whether the judgment that Maxwell's equations are laws has survived over the past hundred years, but rather whether and on what evidence nineteenth-century physicists were justified in expecting these

developments, given the MRL account of what a law is. It is hard to avoid the conclusion that if the MRL account is correct, it is rather miraculous that everything has turned out so nicely. One would expect that whether Maxwell's equations fit "harmoniously" with the other axioms and theorems that make up the best deductive system should be rather sensitive to what those other axioms and theorems claim. According to Earman, despite having very limited information about these other axioms and theorems, and in some cases being completely ignorant of them and of exactly how they fit with Maxwell's equations, scientists were nonetheless able to establish with some confidence that Maxwell's equations were laws. A more natural conclusion is that both what nineteenth-century physicists meant when they claimed that Maxwell's equations express laws and the basis on which they reached this judgment had to do with something else besides their estimate that they were the axioms and theorems of the best deductive system. An invariance-based account is my candidate for this "something else."

## 6.12 Stability, Resilience, and Invariance

As I noted at the beginning of this chapter, a number of other authors have linked laws and causal relationships to notions that bear at least a family resemblance to invariance. In this section, I examine two of these notions, Sandra Mitchell's notion of stability and Brian Skyrms's notion of resiliency, with an eye to clarifying how they differ from invariance.

In a recent series of papers, Sandra Mitchell (1997, 2000) has advocated replacing the standard law-versus-accident dichotomy with a framework for the classification of explanatory generalizations that admits of degrees. One of the "dimensions" of scientific law that receives the most attention within this framework is what Mitchell calls "stability," which is roughly the extent to which a generalization is contingent on conditions that are stable across space and time. Mitchell is particularly interested in the status of biological generalizations like Mendel's "laws": with whether these are genuine laws. She argues that Mendel's laws are less stable (more contingent) than, say, paradigmatic fundamental physical laws like the conservation of mass-energy, but more stable than generalizations like (6.9.5) "All the coins in Clinton's pockets are dimes." Biological generalizations thus occupy an intermediate position on the "continuum of contingency." They are stable enough to play the same role that laws play in other areas of science: they can function so as to represent "causal knowledge" and can be used to predict, explain, and to guide interventions (2000, p. 249). Because biological generalization can function in these ways, we may legitimately describe them as laws, even if they lack features such as exceptionlessness traditionally ascribed to laws.

Mitchell characterizes "degree of stability" as follows:

There is a difference between Mendel's laws and Galileo's law . . . but it is not the difference between a claim that could not have been otherwise



(a “law”) and a contingent claim (a “non-law”). What is required to represent the difference between these two laws is a framework in which to locate different degrees of stability of the conditions on which the relation described is contingent. The conditions upon which the different laws rest may vary with respect to stability in either time or space or both. (p. 252)

Elsewhere, she connects such differences in stability to differences in “the kind of information required to use” different sorts of generalizations (p. 256). Some generalizations, such as the conservation of matter, are contingent on conditions that are relatively stable in space or time. Hence, even in the absence of much additional information, we can with some confidence detach such generalizations from the “evidential context” in which they were originally discovered and apply them to new situations. However, when (as is the case with most biological generalizations) a generalization is contingent on conditions that are much less stable across space and time, “more information is required for application” (p. 257): one cannot assume that the generalization will hold for new situations in the absence of special reasons to think that it does.

I agree, of course, with Mitchell about the desirability of replacing the all-or-nothing law/accident dichotomy with a framework for the classification of explanatory generalizations that admits of degrees. Moreover, it may seem (and Mitchell herself claims) that her notion of stability is closely similar to my notion of invariance (2000, p. 258). I will argue, however, that the notions are rather different in motivation and that neither is necessary nor sufficient for the other.

One way of bringing out the difference between invariance and stability is to return to examples having the structure discussed in section 6.2, in which  $Y$  and  $Z$  are joint effects of the common cause  $X$ . As we have seen, in this sort of case, the relationship between  $Y$  and  $Z$  is not invariant under (any) interventions on  $X$ , hence not invariant at all. However, if the common cause structure itself is stable in the sense that it occurs repeatedly in many or most regions of space and time or in the sense that a single instance of  $X$  has common effects in many spatiotemporal regions, then the relationship between  $Y$  and  $Z$  will be highly stable in Mitchell’s sense. In fact, it is easy to think of examples having this sort of structure; many of them involve cosmological regularities. Consider again

(6.12.1) All regions of space exhibit a uniform 2.7 degrees Kelvin microwave background radiation.

(6.12.1) is certainly highly stable in Mitchell’s sense. However, the explanation for why different regions of space conform to this regularity is that the radiation in these different regions are effects of a single common cause: the conditions that prevailed in the very early universe at the time of the big bang. If we could somehow intervene to alter the microwave background radiation in some particular region of space, this would not be a way of altering the background radiation in other regions. Thus, although (6.12.1) is stable in Mitchell’s sense, it not invariant under (testing) interventions. Because it fails

to be invariant under interventions, (6.12.1) is not a plausible candidate for a law of nature. Consistently with the account of explanation I defended above, it also seems clear that one could not legitimately appeal to (6.12.1) to explain why some particular region of space exhibits the background radiation it does. Instead, the explanation for this is to be found in the conditions prevailing in the early universe. As another example, consider that on an appropriately large scale, the mass distribution of the universe may well be homogeneous. The generalization describing this fact is again highly stable in Mitchell's sense but not invariant, lawful, or explanatory. Stability is one thing; invariance another.

A similar point holds for generalizations in other areas of science. In an interesting recent paper, Kenneth Waters (1998) distinguishes between two sorts of biological generalizations. Generalizations about *distributions* describe "historically based contingencies" concerning the distribution of biological entities or characteristics. Examples include generalizations about the prevalence of various kinds of circulatory systems across taxa or the generalization that major arteries in many organisms contain a larger amount of elastin than other blood vessels. Waters contrasts such generalizations with generalizations expressing *causal regularities*, offering as an example the generalization that blood vessels containing a large amount of elastin will expand when the amount of fluid in them is increased (p. 19). Although Waters does not explicitly endorse a manipulationist account of the content of causal generalizations, his examples fit very naturally into such a framework. Changing the amount of fluid in a blood vessel containing elastin is a way of manipulating whether it expands, and it is because this generalization furnishes information relevant to manipulation that it is appropriate to think of it as a causal or explanatory generalization. By contrast, the generalization that elastin is present in larger amounts in arterial blood vessels in many animals does not describe a relationship that is even potentially relevant to manipulation and control and hence is not a causal or explanatory generalization.

The relevance of this to Mitchell's discussion is as follows. Generalizations describing distributions that claim that some biological characteristic is very widely or universally shared by all organisms may exhibit a great deal of stability in Mitchell's sense (and may be relatively detachable from their original evidential context), but this fact does not show that they are causal or explanatory generalizations. Conversely, a generalization can be causal or explanatory in the sense that it describes a relationship that is exploitable in principle for manipulation and control, even though this relationship holds for or applies only to a biological structure that is not shared by many organisms. As already noted, a generalization describing the response of a particular sort of neural circuit to stimulation can qualify as causal or explanatory even if that circuit is not widely distributed in other organisms.

This difference is in turn connected to another difference between my view and Mitchell's. Mitchell thinks in terms of a single "continuum of contingency," with a generalization like (6.12.2) the law of the conservation of mass-energy at one end and a generalization like (6.9.5) "All the coins in Clinton's pocket

are dimes" at the other (2000, p. 253). This framework seems inevitable if the degree of stability or contingency of a generalization just has to do with the stability across space and time of the conditions on which it is dependent. The difference in stability in this sense between (6.12.2) and (6.9.5) is clearly a matter of degree; assuming that there was a period of time during which (6.9.5) was true, (6.9.5) itself and the conditions on which it depends must have been stable over this spatiotemporal interval, however limited its duration. By contrast, on the account that I advocate, although generalizations can differ in degree of invariance, some generalizations, including (6.9.5), are not invariant at all. These do not differ in degree from invariant generalizations but rather in kind: they fall below the threshold for explanatory status. If, like Mitchell, our interest is in capturing the features a generalization must possess if it is to figure in causal explanations and tell us about the results of interventions, the threshold/continuum model seems more appropriate. It isn't the case that generalizations like (6.12.1) and (6.9.5) are somewhat explanatory but less so than, for example, Maxwell's equations. Rather, (6.9.5) and (6.12.1) seem not to be explanatory and tell us nothing about the results of interventions.

Not only can a generalization be stable in Mitchell's sense without being invariant, but it can also be invariant under some nontrivial range of interventions and changes without being particularly stable in Mitchell's sense. The Hooke's law generalization (6.6.3)  $F = -k_s X$  is (or may be imagined to be) a case in point. (6.6.3) may hold only for a very specialized sort of spring: the particular conditions on which its holding (the fact that the spring has been constructed in a very specific way out of a specific sort of material) depends may occur only very rarely, or within a very small corner of the universe. Nonetheless, when (6.6.3) does hold for some spring, it may be relatively invariant in the sense that it will continue to hold under a fairly wide range of interventions on its extension and under many changes in background conditions (heating the spring may make little difference to its behavior, etc.). As explained above, my view is that when it comes to explaining the force exerted by some particular (sort of) spring  $s$ , what matters is the invariance of (6.6.3) under interventions on this (sort of) spring. How many other springs there are for which (6.6.3) holds and how widely distributed they are in space and time make no difference to how well (6.6.3) explains the behavior of  $s$ .

Similarly for biological generalizations. Consider two biological mechanisms involved in, for example, gene expression and regulation or in brain function. One of these is highly conserved: it is found in many different species of animals that are widely distributed in space and time. The other is specific to a particular kind of animal. A generalization describing the highly conserved mechanism will be more stable in Mitchell's sense than a generalization describing the less conserved mechanism. Nonetheless, both generalizations may be relatively invariant with respect to the behavior of the mechanisms they describe. On my view, the mere fact that a mechanism is highly conserved does not mean that the generalization describing its behavior is more lawful or invariant or that it provides a deeper or better understanding of its behavior

than does a generalization describing a less highly conserved mechanism. In the same way, if certain alleles in some particular population  $P$  conform to Mendel's laws, the fact that there are many or only a few other populations besides  $P$  in which genetic change conforms to Mendel's laws is irrelevant to how well Mendel's laws explain gene change in  $P$ .

These differences between stability in Mitchell's sense and invariance as I conceive it are connected to another difference. Stability in Mitchell's sense appears to be a *de facto*, nonmodal notion. Whether a generalization is stable depends, as Mitchell says, on whether the conditions on which it is contingent are in fact stable across space and time. By contrast, as I have been emphasizing, invariance is a modal or counterfactual notion. Whether a generalization is invariant depends not on whether conditions that would disrupt it in fact occur, but on whether the generalization *would be* disrupted *if* various conditions (involving interventions) *were* to occur. Thus, generalizations like "All the coins in Clinton's pocket are dimes" or the cosmological generalization (6.12.1) about the uniform microwave background fail to be invariant not because conditions that disrupt them do in fact occur, but because *if* certain changes were to occur, they would disrupt them. Mitchell's account threatens to treat any *de facto* regularity as a law, although perhaps one with a very modest degree of stability. This seems to me to elide the crucial distinction between those generalizations that hold pervasively but are entirely accidental and nonexplanatory and those that genuinely represent causal relationships.

I turn next to some remarks on the notion of resiliency in the sense of Skyrms (1980; Skyrms and Lambert 1995) and its relationship to invariance. Several commentators<sup>16</sup> have regarded these notions as closely connected and perhaps virtually identical. My view is that they are different in important respects. Abstracting away from formal details, the resiliency of a proposition has to do with the extent to which its subjective probability remains stable or unchanged (or changes only by some small amount) as one conditionalizes on other truth-functional propositions in some family, all of which are consistent with the original proposition and its denial. Resiliency is thus a measure of degree of epistemic entrenchment or "resistance to belief change" (Skyrms and Lambert 1995, p. 141), in the sense that it indicates the extent to which an agent's degree of belief in a proposition would change under changes in her other beliefs. Laws are just generalizations that are relatively highly resilient (under conditionalization on some appropriate set of other beliefs): "the necessity of laws, like the necessity of causes, is resiliency" (p. 145). As Skyrms and Lambert put it, laws are generalizations that play a central role in our Quinean web of belief in the sense that we would resist giving them up under changes in other beliefs (pp. 139–40).

Whereas resiliency is thus an epistemic or doxastic notion having to do with the relationships among an agent's (or perhaps a scientific community's) beliefs, invariance is a nonepistemic or "objective" notion, the characterization of which has to do with the way the world is, rather than with anyone's beliefs. In particular, invariance has to do with the extent to which a generalization would continue to truly describe the behavior of some system (or

the relationship described by the generalization would continue to hold) under changes that are actual physical manipulations or alterations in the system, rather than under changes in an agent's beliefs or evidence. It follows that a generalization can be invariant even though no one knows that it is, and it can be widely believed that a generalization is highly invariant when in fact it is not. Assigning a generalization a central role in one's web of belief and treating it as resilient does not make it invariant. (Indeed, as we shall see, it is not even tantamount to believing that the generalization is invariant.) Conversely, a generalization can be invariant even though, given an agent's other beliefs and the available evidence, it is highly nonresilient. Consider some newly conjectured candidate for a fundamental law of nature for which there is at present only very weak evidence and which contradicts some apparently well-established scientific claims. Belief in this generalization at present may be relatively nonresilient, but of course this is compatible with the generalization being in fact highly invariant.

Although resiliency is thus relativized to a set of beliefs in a way in which invariance is not, this is not the only difference between the two notions, and resiliency is not just a subjective or doxastically relativized version of invariance. Presumably, the doxastically relativized counterpart to invariance is the notion of *believing* a generalization to be invariant, where to believe that a generalization is invariant is just to believe that it would continue to hold under some class of interventions. Like the notion of resiliency, believing a generalization to be invariant must be defined by reference to an agent's belief state. However, having a highly resilient belief in a generalization is not the same thing as believing it to be invariant. The reason is that, unlike invariance, the notion of resiliency assigns no special significance to stability of belief under changes in other beliefs that have to do with the occurrence of interventions. It is perfectly possible for an agent's degree of belief in some generalization to remain stable under changes in many of her other beliefs  $B_i$  but not under changes in her beliefs that various interventions have occurred. In this case, the agent's belief in the generalization is resilient with respect to the other beliefs  $B_i$ , but the agent does not believe the generalization to be invariant. Conversely, an agent may believe that a generalization would continue to hold under some class of interventions—and hence believe that it is invariant—but would readily give up belief in the generalization if she acquires various other beliefs or if certain kinds of evidence become available.

The significance of this point can be brought out by means of some examples. Consider the generalization

(6.12.3) No human beings live on the other planets of the solar system.

There is a great deal of evidence for (6.12.3), and, in the case of most of us, it is relatively epistemically entrenched in the sense that we would remain committed to it under many possible changes in other beliefs. Nonetheless, despite its resiliency, (6.12.3) is no law of nature. On my view, (6.12.3) is not a plausible candidate for a law of nature because it is not invariant under such

interventions as attempting to establish a colony on Mars. Moreover, it also seems clear that to have a resilient belief in (6.12.3) is not at all tantamount to *believing* that (6.12.3) is invariant or is a law of nature. Most people's degree of belief in (6.12.3) is relatively resilient, but few regard it as a law of nature. The question of whether there are interventions that would render (6.12.3) false seems largely independent of the extent to which (6.12.3) is resilient. A similar remark holds for cosmological generalizations like the generalization (6.12.1) concerning the uniformity of the microwave background. (6.12.1) is supported by a wide variety of evidence, is centrally located in our web of belief, and hence (according to Skyrms and Lambert) relatively resilient, but it does not follow that it is invariant or even that it is believed to be invariant.

Consider another example. My present degree of belief in General Relativity (GR) would change considerably if I became convinced that

(6.12.4) Most experts in the relevant scientific community in 2050 will regard GR as false.

My degree of belief in GR is not resilient under a change in my present beliefs to (6.12.4). Nonetheless, this failure of resilience has nothing to do with the extent to which the equations of GR are invariant. Instead, the issue of invariance has to do with whether there are physical changes or conditions under which those equations would break down. The change consisting in my coming to believe (6.12.4) will change my belief in GR, but this change does not represent a physical change that creates a system in which the equations of GR no longer hold. Similarly, it does not follow from the fact that I would be prepared to give up my belief in GR if I were to become convinced of (6.12.4) (or if various other sorts of evidence were to become available) that I presently believe that the equations of GR are noninvariant or not laws of nature.

An additional point is worth making before leaving the topic of resiliency. This is that many of the examples that Skyrms uses to illustrate the connection between resiliency, on the one hand, and lawfulness and causal necessity, on the other, can be very plausibly interpreted as instead illustrating the connections among invariance, lawfulness, and causation. For example, Skyrms (1980, p. 18) notes (following Max Planck [1960] 1997) that "all attempts to affect" the probability of decay of a uranium atom are unsuccessful. He connects this observation with the claim that generalization (6.12.5) describing this probability of decay is resilient and takes this in turn to capture the sense in which this generalization is regarded by us as lawful or necessary. However, it is at least equally natural to take Planck to be making a claim about the invariance of the generalization (6.12.5): no matter what we or nature do, there are no physically possible changes that will affect the probability of decay. (This is, of course, invariance in the sense of that notion, described in section 6.2, that applies to non-change-relating generalizations.) Planck's language and the particular illustration he offers (that changing the temperature of the atom will not affect the probability of decay) make it clear that he is talking about physical changes in the world, and not (or not just)

about the stability of an observer's belief about the probability of decay under changes in the information or evidence available to him. My suggestion is that it is the invariance of the probability of decay under physical changes that leads us to regard (6.12.5) as lawful or necessary.<sup>17</sup>

In my view, it was an extremely important insight on Skyrms's part to recognize and articulate the connection between the lawfulness of a generalization and whether it remains stable as other conditions are changed. However, the intuitive force of this idea is better captured by thinking of stability in terms of invariance rather than resiliency. Resiliency is an important and valuable concept in its own right, but its proper role is in capturing ideas having to do with doxastic entrenchment and stability of belief under additional information rather than in capturing notions like "law" and "physical necessity."

### 6.13 Explanation and Invariance in Biology

The question of whether there are laws in biology and the role that biological generalizations like Mendel's laws play in evolutionary explanations has attracted substantial attention from philosophers of biology. My aim in this section is to show how the ideas about explanation and invariance developed above can cast light on these issues.

In a very interesting series of papers, John Beatty has advanced what he calls the Evolutionary Contingency Thesis, according to which "all generalization about the living world are [either] just mathematical, physical or chemical generalizations or [if they are] distinctively biological describe contingent outcomes of evolution" (1995, pp. 46–47). He takes this to mean that "there are no laws of biology. For whatever laws are, they are supposed to be more than just contingently true" (p. 46). Beatty illustrates this thesis by means of a discussion of Mendel's first "law" (or law of segregation). He takes this to claim that

(S) with respect to each pair of genes in a sexual organism, 50% of the organism's gametes will carry one representative of that pair and 50% will carry the other representative of that pair. (pp. 50–51)

Beatty draws attention to two related features of this law. First, it has a number of "exceptions." One of the best known involves meiotic drive, which occurs when an allele influences meiosis in such a way that it has a greater than 50 percent chance of ending up in a gamete, rather than the 50 percent chance that Mendelian segregation would require. When Mendel's law is understood along the lines of (S), this phenomenon represents a genuine violation of the law (i.e., a case in which the antecedent but not the consequent of the law holds) and not a mere failure of the law to apply in the sense that its antecedent is not satisfied. Second, the widespread prevalence of Mendelian segregation is itself the result of the operation of natural selection. That is, if, as appears to be the case, most gene pairs in sexual organisms conform to (S), this is because natural selection has operated in such a way as to produce this outcome:

because segregation in accord with (*S*) conferred a selective advantage of some kind. If the past histories of most organisms had been sufficiently different, and if they had been subject to sufficiently different selective forces, nature would have contained few if any genes that segregate according to Mendelian ratios. Thus, whether violations of Mendel's laws occur at all and whether they are common or rare is contingent on the course of evolution. To express the point in the language of this chapter, (*S*) fails to be invariant under many possible changes in selection pressure. This does not mean just that there are at present organisms and populations in which meiosis fails to conform to (*S*), but rather that, even for those types of organisms that presently conform to (*S*), there are possible changes, due to natural selection, that would make it the case that those organisms or their descendants would violate (*S*).

As noted above, Beatty infers from this that Mendel's law is not "necessary" and hence not a real law. Beatty does not explore the implications of this claim for explanatory practice in biology, but a number of other philosophers recently have,<sup>18</sup> and their discussion leads straightforwardly to the dilemma described in section 6.1. If explanation in biology requires appeal to laws and if Beatty is correct about the nonlawful status of (*S*), it appears to follow that the numerous evolutionary models that appeal to (*S*) are unexplanatory, assuming, as appears to be the case, that no other generalizations in these models qualify as laws. Indeed, assuming that Beatty is right in claiming that other distinctively biological generalizations also will be contingent on the course of evolution, it follows that all distinctively biological theory is unexplanatory. Because most, although by no means all, writers regard this as a patently unacceptable conclusion, the most common response has been to search for a somewhat weakened or watered down notion of law, according to which generalizations like (*S*) may qualify as laws and hence as explanatory.

One of the central claims of this chapter is that this response is unnecessary and that it simply distracts from understanding how explanations that appeal to (*S*) work. In assessing the explanatory import of (*S*), what matters is not whether we decide to bestow on it the honorific "law," but the range of what-if-things-had-been-different questions it can be used to answer and the range of changes over which it is invariant. As long as (*S*) is invariant in the right way, it doesn't matter whether it has exceptions or is contingent on the course of evolution in the way that Beatty describes: it still can be used to explain.

To explore this issue, consider some of the ways in which (*S*) figures in elementary explanatory evolutionary models. First, from (*S*) and the assumption of random mating, one can derive the so-called Hardy–Weinberg law, which tells us that in the absence of various evolutionary forces—mutation, migration, drift, and selection—genotypic frequencies will reach equilibrium after one generation, with the equilibrium frequencies depending on the allele frequencies with which we began. One can then use the Hardy–Weinberg law in conjunction with additional assumptions about differential fitness of various genotypes to explain how the frequencies of those genotypes will change in



response to natural selection. Consider a single locus model in which the  $A$  allele is dominant and the  $a$  allele recessive, with the  $Aa$  genotype identical in phenotype and fitness to the  $AA$  genotype and both superior in fitness to the  $aa$  genotype. By using ( $S$ ) and the other assumptions of this model, one can readily derive that after one generation, the frequency of  $a$  will decrease by an amount that is a function of its initial frequency and its relative fitness. In particular, the change  $\Delta q$  in allele frequency of  $a$  will be:

$$(6.13.1) \quad \Delta q = q(1 - wq^2) / (1 - wq^2) - (q - wq^2) / (1 - wq^2) = -wpq^2 / (1 - wq^2)$$

where  $p$  is the initial frequency of the  $A$  allele,  $q$  the frequency of the  $a$  allele, and  $w$  the relative fitness of the recessive homozygote. With no changes in relative frequencies, this process will continue in subsequent generations until the  $a$  allele is eliminated. By contrast, in the frequently discussed case of heterozygote superiority, in which the heterozygote  $Aa$  is more fit than either of the homozygotes  $AA$  and  $aa$ , one can show that selection will lead to a polymorphic equilibrium in which both the  $A$  and  $a$  alleles are maintained in the population, the frequency of each being a function of the fitnesses of the various genotypes.

Models and derivations of this sort are commonly regarded as explaining why gene frequencies change or fail to change over time, and the account of explanation developed in previous chapters supports this judgment. Just as with the derivations described in section 5.1, we can think of these derivations involving ( $S$ ) as explanatory in virtue of locating various potential explananda within a range of alternative possibilities and answering a set of what-if-things-had-been-different questions about these explananda. In consequence, we are given in each case a sense for the factors or conditions on which these explananda depend. For example, given the assumptions that figure in the derivation of the Hardy–Weinberg law, one can see that changing the values of the initial allelic frequencies will not change whether an equilibrium is reached, but will change the genotype equilibrium frequencies; we can see how these frequencies would have been different if the initial allelic frequencies had been different. We can also see from this derivation how matters would be different in populations that do not conform to Mendelian segregation. In such populations, even in the absence of migration, selection, and so on, the Hardy–Weinberg law will not hold. Instead, one form of the gene will completely replace the other form in the population.

Similarly, in the case in which there is selection operating at a single locus against a recessive homozygote, one can see how the outcome would (or would not) have been different in various ways if the initial frequencies  $p$  and  $q$  of the alleles  $A$  and  $a$  and the relative fitness  $w$  of the recessive homozygote had been different. In particular, the derivation shows us that, provided selection is allowed to run on long enough and no counteracting forces are operative, the recessive homozygote will always be eliminated regardless of the particular values of  $p$ ,  $q$ , and  $w$ . By contrast, the *rate* at which the allelic frequencies change in response to selection depends on the exact value of the selection coefficient  $w$  and the initial allelic frequencies. In this case, as well as

in the previous examples, the model shows us how the change  $\Delta q$  in allelic frequency per generation would change if the value of  $w$  or  $q$  were changed, that is, how  $\Delta q$  would have been different had such changes occurred.

Parallel remarks apply to case of heterozygote dominance. In this case, the model described above allows us to see that the maintenance of polymorphic equilibrium is directly attributable to the superior fitness of the heterozygote. If environmental circumstances were to change in such a way that the heterozygote was no longer superior in fitness—as presumably has happened in the case of the sickle cell anemia in those areas from which malaria has been eradicated—then, in the absence of countervailing forces, the recessive allele would be eliminated from the population. Similarly, the model allows us to see how, when the heterozygote is superior in fitness, the equilibrium frequency with which it would be maintained in the population would be different in various ways, depending on the relative fitness of the heterozygote in comparison with the two homozygotes.

What must be true of the generalization ( $S$ ) if it is to contribute to answering a range of what-if-things-had-been-different questions in the way just described? According to the position defended in this chapter, ( $S$ ) must be stable or invariant in the right way in the population  $P$  whose behavior we are trying to explain. This requires that ( $S$ ) must hold, or hold approximately, in the populations  $P$  whose behavior (i.e., changes in genotype frequencies) we are trying to explain, but it also requires that ( $S$ ) continue to hold under certain changes in those populations. In particular, because the general strategy in the above models is to use ( $S$ ) in conjunction with information about initial genotype frequencies and the fitness of various genotypes to explain patterns of changes in those frequencies, ( $S$ ) must be invariant under changes (including changes produced by interventions) in those frequencies and fitnesses in population  $P$ . That is, it should be true that changes in, say, the frequency of the  $Aa$  genotype or in its relative fitness in  $P$  do not by themselves disrupt (at least over the time scale of changes in gene frequency we are trying to explain) the mechanism of normal Mendelian segregation in  $P$ . We know that this condition is met for many populations conforming to ( $S$ ). The mere fact that a genotype becomes more or less fit as a consequence of some environmental change or that it increases or decreases in frequency will not by itself cause a shift from Mendelian to non-Mendelian patterns of segregation.

In addition, as we have seen, satisfaction of the condition that ( $S$ ) be invariant under some range of interventions that change gene frequencies and fitnesses in  $P$  is quite compatible with there being widespread exceptions to  $S$  or  $S$ 's failing to be invariant in *other* populations. If ( $S$ ) is not invariant in some other population  $P^*$ , we cannot appeal to ( $S$ ) to explain why genotype frequencies change as they do in  $P^*$ , but this, by itself, does not undermine the use of ( $S$ ) to explain changes in  $P$ . Moreover, for reasons that now should be clear, the invariance of ( $S$ ) under changes in gene frequencies and fitnesses in  $P$  is perfectly compatible with ( $S$ ) failing to be invariant under the (other) sorts of changes in selective regime described by Beatty—whether these occur in  $P$  or in some other population. On the account of explanation I have been defending,

what matters for the successful use of (*S*) in the explanations described above is not that it be invariant under all possible changes, but merely that it be invariant in a much more limited way: under changes in genotype frequencies and relative fitnesses that will occur in *P* over a limited time period. Thus, even if we agree with Beatty that (*S*) is not a real law because there are evolutionary changes over which it fails to be invariant, we can still legitimately appeal to it to explain in circumstances of the sort described above.

This example illustrates a more general point worth underscoring. Whether a trait conforms to Mendelian segregation is contingent on what selective pressures happen to be present in the organism's environment. Whether (6.2.6) "All the coins in Clinton's pockets are dimes" holds is contingent on such factors as the decisions of Clinton and others. However, there is an important difference between the sort of contingency present in these two cases. The contrast between (*S*) and (6.2.6) shows that whether a generalization can be used to explain depends not on whether or not it is contingent on some additional condition, *but rather on the details of the way it is so contingent: on just what the condition in question is and on how it is related to what we are trying to explain.* If we merely ask whether various biological generalizations are contingent, we miss important differences between generalizations like (*S*) and generalizations like (6.2.6). It is crucial to understanding these differences that the condition on which (*S*) is contingent—whether Mendelian segregation continues to provide a selective advantage—is itself not changed by the kinds of changes in fitness or initial genotype frequency that we invoke when we appeal to (*S*) to explain. The corresponding claim is not true of (6.2.6): a testing intervention that changes which coins are in Clinton's pocket will itself change or disrupt the condition on which the truth of (6.2.6) is contingent. For example, such an intervention would require that any policy of Clinton of allowing only dimes into his pocket no longer be effective.

This example also illustrates the claims made previously about what one must know in order to explain (section 6.8). For the models described above to provide understanding, one must know (*S*) itself and that it holds invariantly over some appropriate range of changes in the populations whose behavior one is trying to explain: one must know that these populations are within the domain of invariance of (*S*). However, it is not required that one be able to provide a precise and informative specification of the exact boundaries of the domain over which (*S*) holds. That is, one does not have to be able to provide or exhibit an exceptionless generalization (*S*<sup>\*</sup>) along the lines of "when circumstances *C*<sub>1</sub> . . . *C*<sub>*n*</sub> obtain, then (*S*) holds." I take this to correspond to actual practice in evolutionary biology. As we have already noted, there are a number of different circumstances in which (*S*) breaks down. Biologists who appeal to Mendelian segregation to explain do not concern themselves with formulating exceptionless generalizations along the lines of (*S*<sup>\*</sup>); indeed, they probably believe that they are in no position to do this. Instead, as the passage quoted from Dunn in section 6.8 illustrates, they appeal just to (*S*) itself and to various sorts of informal empirical arguments designed to make it plausible that (*S*) holds invariantly for the populations they are trying to model. Again,

the model of explanation that I have proposed makes much better sense of this aspect of biological practice than traditional nomothetic models.

These claims about the explanatory role of (S) also illustrate my earlier remarks about the difference between other-object and same-object counterfactuals. Assuming again that our goal is to explain gene frequencies at equilibrium or rates of genic change in some particular population *P*, same-object counterfactuals have to do with, for example, how gene frequencies *in P* would change if initial gene frequencies or fitnesses *in P* were to be changed by interventions: that is, with whether generalizations like (S) would continue to correctly describe what would happen in *P* under such interventions. By contrast, other-object counterfactuals have to do with whether various *other* populations, different from *P*, conform to Mendel's laws or with the conditions under which such populations would conform to Mendel's laws. My account insists that it is same-object counterfactuals that are crucial to successful explanation. The idea that same-object counterfactuals are the ones that matter for explanation is closely tied to my previous claim that the use of Mendel's generalizations to explain in connection with populations that exhibit Mendelian segregation is not undercut by the existence of other populations for which those generalizations fail to hold and is also not undercut by our inability to formulate some noncircular general condition that distinguishes between these two kinds of populations.

#### 6.14 Invariance and Ceteris Paribus Laws

It will help to bring out what is distinctive in the ideas about invariance and explanation that I have been defending if we contrast them with a standard alternative account of the conditions that explanatory generalizations in the special sciences must meet. I call this the *completer* account; versions can be found in many writers, from Hempel (1965b), to Fodor (1991), Hausman (1992), and Pietroski and Rey (1995). Each of these writers adds refinements and complications, but I will focus on the core idea, arguing that it is sufficiently mistaken that the embellishments won't help.<sup>19</sup> The completer account adopts the nomothetic assumption that all explanation requires "subsumption" under laws, and also assumes that a necessary condition for a generalization to count as a "strict" or unproblematic law is that it be exceptionless (recall the passage from Pietroski and Rey 1995 quoted in section 6.7). We then face the familiar problem of reconciling these assumptions with the apparent paucity of exceptionless generalizations in the special sciences. In schematic form, the solution proposed by the completer account is this: suppose one begins with a generalization of the form

(6.14.1) All *F*s are *G*s

which has exceptions. (6.14.1) will be a legitimate kind of law, a so-called *ceteris paribus* law, and will have explanatory import if and only if there is some further condition *C* such that

(6.14.2) All *F*s in *C* are *G*s

is a strict or exceptionless law and neither *F* by itself nor *C* by itself is nomologically sufficient for *G*. Adopting (and somewhat modifying) some terminology due to Fodor (1991), let us call such a condition *C* a “completer” for (6.14.1). The simplest version of the completer account then says that a *ceteris paribus* generalization is genuine law and hence explanatory if and only if it has a completer. It is crucial to the structure of this account that we *not* impose the requirement that someone who appeals to (6.14.1) to explain must be able to actually describe its completer *C* in a nontrivial way or to state the exceptionless generalization (6.14.2). As I have repeatedly noted, it is rarely possible to do this in the special sciences. Instead, it is enough that the completer exists or, alternatively, perhaps that we know or reasonably believe that it exists, even if we are unable to provide a nontrivial description of it.<sup>20</sup>

The apparent attraction of the completer strategy is that it allows one to retain the idea that there is a sense in which explanation requires laws and that laws must be exceptionless, while at the same time according an explanatory role to generalizations that have exceptions; this is accomplished by requiring that (6.14.1) be “backed” or associated with the exceptionless law (6.14.2). The idea is that somehow (6.14.2), in virtue of its exceptionlessness, endows (6.14.1) with explanatory import and gives it a status as a legitimate (*ceteris paribus*) law that it would not possess if it did not have a completer. We can also think of this strategy as illustrating a general point made in section 6.10: that the claim that a generalization like (6.14.1) is a “law” (albeit a *ceteris paribus* law) becomes a substantive claim and not merely a recommendation about terminology when it is embedded in a more general set of ideas about the features a generalization must possess (in this case, completability into an exceptionless generalization) if it is to be explanatory.

Despite its apparent naturalness, I believe that the completer strategy is fundamentally flawed and that an appreciation of these flaws will bring out the superiority of the invariance-based account. First, there is a familiar epistemic puzzle. Consider the simplest version of the completer account, according to which the mere existence of a completer endows a generalization with explanatory import even if those who use the generalization are unaware of its existence. It is far from obvious how this is possible. As argued in chapter 4, it is a plausible principle that explanatory information must be information that is epistemically accessible: explanations work by conveying information that provides understanding, and this means that such information must be epistemically accessible to those who provide the explanation or are enlightened by it. If some of the information allegedly conveyed by an explanation cannot be grasped or recognized by those who use the explanation, it is not through recognition of that information that the explanation provides understanding. Thus, if no one knows that (6.14.1) has a completer (or even what a completer is), and if the structure of what is known cannot be represented by (6.14.2), yet (6.14.1) is used to explain, it is hard to see how its explanatory import can be crucially bound up with whether it has a completer. A parallel assessment is

warranted if it is instead claimed that for (6.14.1) to be explanatory those who appeal to it or those in the audience to which it is directed must know (or truly) believe that it has a completer, even if they do not know (or have true beliefs about) the identity of the completer. What is the theory of explanation according to which merely knowing that (6.14.1) has a completer but not the identity of the completer endows it with explanatory import?

By contrast, on my account, explanatory import depends only on epistemically accessible information. Those who appeal to explanatory generalizations that are not exceptionless laws typically are able to recognize that they are within the domain of invariance of those generalizations and are able to see how they can be used to answer a range of what-if-things-had-been-different questions. Information that is not epistemically accessible to users, such as information about the exact boundaries of domains of invariance or the full range of circumstances in which a generalization will break down, is not information that is needed to successfully explain.

A second problem is that the underlying motivation for the completer account is problematic. This motivation depends on the idea that there is an invidious contrast between *ceteris paribus* laws like (6.14.1), which have exceptions, and genuine or strict laws that are exceptionless, and that, because of this, to vindicate the former we must show that they are backed in an appropriate way by the latter. However, as we have already observed and as defenders of the completer account readily acknowledge, there are relatively few examples of exceptionless laws to be found anywhere in science, even in physics. (Indeed, both Fodor 1991 and Pietroski and Rey 1995 explicitly suggest that there may be *no* known examples of exceptionless laws.) Surely, the natural conclusion is that the whole idea that genuine laws must be exceptionless and that explanation requires exceptionless laws needs rethinking. It is just this conclusion that I have advocated in this chapter.

The third and most fundamental problem is that the basic intuition underlying the completer strategy is wrong: the distinction between those generalizations that have completers and those that do not does not coincide with the distinction between those generalizations that are lawful (or invariant or explanatory) and those that are not. Under the assumption of macrodeterminism, which virtually all defenders of the completer strategy endorse, there are many generalizations that have completers that no one would regard as explanatory or as *ceteris paribus* laws. Consider the generalization

(6.14.3) All human beings with normal neurophysiological equipment speak English with a southern U.S. accent.

This generalization is of course false—it has exceptions—but under the assumption of determinism there will be a very complicated set of conditions *K* that are nomologically sufficient in conjunction with being a human being with a normal neurophysiology for speaking English with a southern accent and that satisfy the other conditions for being a completer, such as those described in Pietroski and Rey (1995). Indeed, we even have a general sense of what those conditions are: they include some very complex set of environmental

conditions, including appropriate early exposure to English spoken with a southern accent. These, together with being a human being with the appropriate neurophysiological structures, are nomologically sufficient to ensure that one will learn to speak English with a southern accent. So (6.14.3) has a completer:

(6.14.4) All human beings with normal neurophysiology in *K* speak English with a southern accent.

Moreover, (6.14.4) is not just exceptionless, but arguably satisfies many of the other standard conditions for lawfulness, such as support for counterfactuals. Nonetheless, (6.14.3) is surely not a generalization that anyone would regard as a *ceteris paribus* law. For one thing, parallel reasoning will also establish that the generalizations “All human beings speak Chinese (or Hindi or Spanish)” have completers and hence are also *ceteris paribus* laws. Nor is the problem that (6.14.3) has exceptions. Even if we imagine that (6.14.3) is exceptionless—that, as the result of political and economic changes, all living humans were to come to speak English with a southern accent, and even if past history had been different in such a way that throughout all history, human beings spoke English with a southern accent—(6.14.3) would not be a plausible candidate for a law of any sort, *ceteris paribus* or otherwise. Nor can we appeal to (6.14.3) to provide an explanation of why some particular person speaks English with a southern accent.<sup>21</sup> (We can, of course, appeal to (6.14.4) to explain this, but (6.14.4) is not (6.14.3).)

By contrast, the invariance-based account does explain in a natural way why (6.14.3) is a poor candidate for an explanatory generalization: it is either noninvariant or invariant only under a very narrow range of interventions. Even if, as a result of political changes, it becomes true that everyone in the world speaks English with a southern accent and even if, because the past history of the human race was very different, all human beings in the past had spoken southern English, (6.13.3) would still be highly fragile. Its truth would depend on a great many very specific contingencies, and if these were to change, (6.14.3) would be disrupted. Only in a very special and rare kind of environment (rare in comparison with the full range of environments in which human beings learn to speak some language or other) do human beings learn to speak English with a southern accent.

We can further bring out the difference between the completer account and the invariance-based account by considering what the former has to say about the contrast between, on the one hand, the “shallow” generalization (6.4.1) linking the position of the gas pedal and the speed of a car and, on the other, the deeper engineering style theory (6.4.2) described in Haavelmo’s example (section 6.4). (6.4.1) has exceptions but, according to the completer strategy, will qualify as a *ceteris paribus* law because there exists some complicated condition *K* (specifying the details of the functioning of the car engine and the environmental conditions in which it is operated) such that the generalization “In *K*, (6.4.1) holds” is an exceptionless law. For similar reasons, (6.4.2) will also qualify as a nonstrict, *ceteris paribus* law: there will

be circumstances in which it breaks down, but it will also have a completer. However, the invariance-based approach allows us to say that (6.4.2) furnishes a deeper explanation of the behavior of the car because (6.4.2) is invariant under a wider range of interventions than (6.4.1) and can be used to answer a wider range of what-if-things-had-been-different questions. By contrast, the completer strategy provides no basis for such a discrimination; all it says about (6.4.1) and (6.4.2) is that both are *ceteris paribus* laws. Again, I take this to illustrate how the invariance-based account focuses on a very different set of considerations in assessing the explanatory credentials of a generalization than the completer strategy. Simply asking whether a generalization has a completer gives us no insight into the range of changes over which it is invariant and the range of *w-questions* it can be used to answer. Yet, it is just these questions that are crucial to explanatory assessment.

Suppose that we are given a generalization of the form (6.14.1) “All *Fs* are *Gs*,” which has exceptions but also has a completer *C*. The range of invariance of (6.14.1) will depend not just on whether the completer *C* exists, but on the details of the way the holding of (6.14.1) depends on both on *C* and on the various alternatives to *C*: on whether, for example, *C* represents a special case, with (6.14.1) holding only when *C* does, or whether, on the other hand, *C* is a more generic case and (6.14.1) would continue to hold under some range of alternatives to *C*. Again, the relevance of these sorts of considerations to explanatory assessment are lost if we focus only on the question of whether (6.14.1) has a completer.

## 6.15 Invariance and Possible-Cause Generalizations

Aside from my remarks about the completer strategy, my characterization of invariance has largely focused on relationships that are *formulated* as (or purport to be) exceptionless, even though we may have good reason to suppose that in fact they are far from exceptionless—relationships such as the ideal gas law and the Mendelian law of segregation. What about other sorts of relationships? In particular, can the notion of invariance be extended to qualitative possible-cause relationships of the sort described in section 5.8, relationships such as “Smoking causes lung cancer” and “Untreated syphilis causes paresis” that are rough and gappy and do not even purport to be exceptionless? I believe that it can. The basic idea, which is foreshadowed in 5.8, is that a possible-cause generalization “*Cs* cause *Es*” will be relatively invariant to the extent that *Es* occur more frequently when *Cs* are introduced by interventions than when *Cs* are absent across a wide or large range of changes in circumstances or background conditions, although not necessarily in *all* such circumstances. We may have good reason to believe that “*Cs* cause *Es*” is invariant in the weak exception-permitting sense just described, even though we do not know any deterministic or probabilistic *law* (in what I have called the “strong” sense of law) linking *Cs* to *Es*. That is, we might have compelling evidence that *Es* occur more frequently when *Cs* are introduced by



interventions than when Cs are absent in a wide variety of circumstances, even though there is no known generalization of the form Cs in *K* are always followed by Es that we are prepared to regard as a law and even though the amount by which the introduction of Cs raises the probability of Es varies in an irregular way, even when other known relevant variables are controlled for. In such a case, what is stable under interventions across different contexts is simply the fact that there is some correlation or other between Cs and Es.

As argued elsewhere (Woodward 1993a), we are most comfortable saying that Cs have the *capacity* or *power* to produce Es when the possible-cause generalization linking Cs to Es is relatively invariant in the sense just described. To speak of a capacity in this sense is to imply that the ability of Cs to cause some characteristic effect *E* is something that, in Nancy Cartwright's words (1989a), *C* "carries around with it" from situation to situation. To return to an example from chapter 5, although one can certainly imagine circumstances in which particular noteworthy events cause fires, they presumably do so only under very special and unusual conditions. Although it is also true that short circuits will cause fires only in appropriate background circumstances, there is a considerable range of different circumstances in which this is true, and it is this fact that leads us to think of short circuits but not noteworthy events as having the capacity to cause fires.

A second, more "scientific" illustration of a possible-cause generalization that is relatively invariant is provided by a well-known paper by Cornfield, Haenszel, and Hammond (1959), which describes evidence that, as of 1957, was regarded by the authors as convincingly showing that smoking causes lung cancer. This paper was written in the absence of a detailed knowledge of the precise biochemical mechanism by which smoking produces lung cancer, and relied largely on epidemiological evidence and to a lesser degree on experimental studies of laboratory animals. Cornfield et al. make no attempt to formulate a law or exceptionless generalization describing the conditions under which smoking will be followed by lung cancer or to formulate a law specifying the probability that one will develop lung cancer given that one smokes. Instead, they lay a great emphasis on the relative stability or invariance of the relationship between smoking and lung cancer. For example, the authors note that some association appears between smoking and lung cancer in every well-designed study on sufficiently large and representative populations with which they are familiar. There is evidence of a higher frequency of lung cancer among smokers than among nonsmokers, when potentially confounding variables are controlled for, among both men and women, among people of different genetic backgrounds, across different diets, different environments, and different socioeconomic conditions (p. 181). The precise level and quantitative details of the association do vary, for example, the incidence of lung cancer among smokers is higher in lower socioeconomic groups, but the fact that there is some association or other is stable or robust across a wide variety or different groups and background circumstances.

A similar stable association is also found among laboratory animals exposed to tobacco smoke. In addition, generalizations about the "influence"

(in the sense described in section 5.8) of smoking on lung cancer and in particular generalizations having to do with the manner or mechanism or *modus operandi* by which smoking produces lung cancer are similarly stable across changes in circumstances in just the way we would expect if smoking causes lung cancer: heavy smokers and inhalers in all populations show a higher incidence of lung cancer than light smokers and noninhalers, groups that have smoked the longest show the highest incidence of lung cancer, and the incidence of lung cancer is lower among those who have smoked and stopped than among relevantly similar groups who are still smoking.

Thus, although Cornfield et al. do not exhibit a precise deterministic or probabilistic generalization that is invariant across different circumstances or populations, the cumulative impact of their evidence is to show that the relationship between smoking and lung cancer is relatively invariant in the weak sense described above. This is enough to show that smoking has the capacity to cause lung cancer and that one may appeal to smoking behavior to explain the occurrence of lung cancer on both the individual and group level.

It may seem that this loose, exception-permitting notion of invariance is so permissive that it excludes almost nothing. This is not the case. First and most obvious, generalizations like that describing the correlation between barometer readings and storms will not be invariant even in the weak sense just described. A more interesting possibility is described by Michael Oakeshott (1966). Oakeshott claims that “nothing in the world of history is negative or non-contributory” (p. 208). His picture is one in which all historical events are “connected” or “contribute” to one another in a seamless web of reciprocal interdependence (pp. 207ff). One way of understanding these remarks is as follows: if, say, the French Revolution had not occurred, then everything that followed this event and the entire web of relationships in which it is embedded would have been different in some way, but there are no grounds for even an educated guess (and it is no part of the task of the historian to speculate) about the specific determinate way in which subsequent history would have been different. Moreover, it is not just that we cannot tell which counterfactuals with this antecedent and specific determinate consequents are true; rather, such counterfactuals lack a determinate truthvalue. Similarly, counterfactual claims about how, if the French Revolution had occurred in a different historical context, subsequent developments would have been different in various specific respects (e.g., “If the French Revolution had occurred thirty years later, Hitler would not have come to power”), will lack truth values. Oakeshott claims that in a world (or domain of inquiry) in which such interdependence is present, the notions of causation and causal explanation would be inapplicable. Whether or not we accept this conclusion, it is clear that in such a world something that we take to be important to causal understanding would be lacking. Oakeshott’s historical world is not merely one in which historical events are not law-governed; it is a world in which historical events fail to have characteristic effects at all to which they are linked in a way that satisfies even the weak invariance condition described above. In such a world, there are no causal capacities that remain stable across different contexts.

My guess is that Oakeshott's claims about what the "world" of history is like are more plausible for some areas of historical inquiry than others. In economic history, it is plausible that there are many possible-cause generalizations that are at least weakly invariant. By contrast, many relationships in intellectual and cultural history may well approximate the kind of seamless interdependence that Oakeshott describes. It is striking that these are exactly areas in which historians are most likely to argue that the search for causal understanding is inappropriate or misguided and ought to be replaced by some other goal ("interpretive plausibility," "hermeneutic understanding," "thick description," etc.). Whether or not we agree with this assessment, on the account defended above it will be a broadly empirical matter whether the relationships among the objects, events, or processes in some particular domain of inquiry will satisfy the weak invariance condition; there is no a priori guarantee that every subject matter must be a suitable domain for causal explanation.

## Causal Interpretation in Structural Models

Although regression equations and more complex structural equation models are widely used in the biomedical, behavioral, and social sciences to represent causal relationships, they have received surprisingly little philosophical attention. My aim in this chapter is to use the ideas about interventions and invariance developed in previous chapters to illuminate the structure of such models.<sup>1</sup>

### 7.1 Regression

I assume that many readers are unfamiliar with structural equation models and hence that some brief orientation will be useful. I begin with simple linear regression involving just one independent variable. Suppose that  $X$  and  $Y$  are variables, measurable on an interval scale, and that we have  $n$  pairs of observations  $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$  of  $X$  and  $Y$  for the individuals in some population  $P$ . Interpreted literally, a linear regression model claims that, except for the operation of a so-called error or disturbance term,  $U$ , there is a definite general linear relationship between  $X$  and  $Y$ , that is, that

$$(7.1.1) \quad y_i = A + Bx_i + u_i \quad \text{for } i = 1 \dots n$$

where  $A$  and  $B$  are fixed coefficients that characterize *all* individuals in the population and  $u_i$  is the value of  $U$  for the  $i$ th observation, where  $U$  is assumed to vary for different observations. It is usual to think of  $U$  as reflecting either “measurement error” in  $Y$  (but not in  $X$ ) or as resulting from the operation of various other variables besides  $X$  that causally influence  $Y$  but have been left out of (7.1.1). Because my interest is in causal interpretation, I focus mainly on this latter possibility. One assumes that the  $u_i$  are values of a random variable  $U$  with a definite probability distribution and hence that  $Y$  is a random variable as well. The operation of the disturbance term  $U$  will thus result in a “spread” of values for  $Y$  for fixed values of  $X$ . In general, the regression equation for  $Y$  on  $X$  is the equation that gives the path of the mean value of  $Y$  for fixed values of  $X$ . In the specific context of linear regression, the problem of specifying the regression equation will reduce to the problem of estimating the values of the

parameters  $A$  and  $B$  in (7.1.1). To do this requires the choice of an estimating procedure and certain assumptions about the distribution of the disturbance term. The usual practice is to employ the method of "least squares," that is, to choose estimators  $A^*$  for  $A$  and  $B^*$  for  $B$  such that the quantity

$$(7.1.2) \quad Q = \sum(y_i - A^* - B^*x_i)^2$$

is minimized. Geometrically, this corresponds to choosing the regression line so that it "best fits" the scatter of points  $(x_i, y_i)$ , where the criterion of best fit is the minimization of the squared vertical distance of these points from the regression line. The ordinary least squares (OLS) estimators obtained by minimizing  $Q$  in (7.1.2) have desirable properties if the following assumptions about  $U$  hold.

(7.1.3) (a)  $E(U_i) = 0$ , (b) homoscedasticity or constancy of the variance of  $U$  across different values of  $X$ :  $V(U_i) = \sigma^2$ , (c) statistical independence or absence of auto-correlation in the error term, (d) statistical independence of the error term and the independent variable; that is,  $X$  and  $U$  are independent.

Under assumptions (7.1.3a–d), one can show that the least squares estimators  $A^*$  and  $B^*$  of  $A$  and  $B$  are best linear unbiased estimators.<sup>2</sup>

We may obtain the conditions under which  $Q$  in (7.1.2) is minimized by taking the partial derivatives of  $Q$  with respect to  $A^*$  and  $B^*$  and setting these equal to 0. The result is that the least squares estimators for  $A$  and  $B$  are

$$(7.1.4) \quad B^* = S_{xy}/S_x^2$$

$$(7.1.5) \quad A^* = \bar{Y} - B^*\bar{X}$$

Here,  $S_{xy}$  and  $S_x$  and  $S_y$  are, respectively, the sample covariance between  $X$  and  $Y$  and the standard deviations of  $X$  and  $Y$ , and  $\bar{X}$  and  $\bar{Y}$  are the sample means for  $X$  and  $Y$ .

Thus, given certain assumptions about the functional relationship between  $X$  and  $Y$  and the distribution of the error term, one can derive estimates for the regression coefficients  $A$  and  $B$  in (7.1.1) from facts about the observed or directly measured values of  $X$  and  $Y$  (because  $S_{xy}$ ,  $S_x$ , and  $S_y$  can be inferred from the data). The result will be a linear equation relating  $X$  and  $Y$ . For example, in his introductory econometrics textbook, Maddala (1977) regresses a variable  $C$  representing expenditures per capita (1958 prices) on a variable  $Y$  representing disposable income per capita (also 1958 prices) for values of  $C$  and  $Y$  for the years 1929–1970 and obtains the following linear equation:

$$(7.1.6) \quad C = 55.432 + .8735Y$$

This suggests that an average increase of \$.87 in per capita consumer spending will be associated with each dollar increase in disposable income.

(Here, “associated” means just that; nothing has been said so far about what would justify us in assuming that changes in  $X$  *cause* changes in  $Y$  in accord with (7.1.6). This topic is addressed below.)

Multiple regression involves a generalization of this model to the relationship between a dependent variable  $Y$  and a number of independent variables  $X_1, X_2 \dots X_k$ . In the linear case, one assumes the model

$$(7.1.7) \quad y_i = B_0 + B_1x_{1i} + B_2x_{2i} + \dots B_kx_{ki} + u_i, \quad i = 1, 2 \dots n$$

where the  $y_i$  are the  $n$  observations one makes on  $Y$  and  $x_{1i} \dots x_{ki}$  are the corresponding observations one makes on the variables  $X_1 \dots X_k$ , and  $u_i$  are, as before, the values of a disturbance term  $U$ . It will be convenient to write (7.1.7) in matrix notation as in figure 7.1.1. (The role of the column of 1s added to the  $X$  matrix is to provide constant multipliers for  $B_0$  and  $B'$  is just the transpose of  $B$ .)

Using again the least squares criterion of fit, one can show that if one makes assumptions like (7.1.3a–d) regarding the disturbance term  $U$ , the best linear unbiased estimator for the vector  $B$  is, in close analogy with (7.1.4), the vector

$$(7.1.8) \quad B^* = (X'X)^{-1}X'Y$$

where  $X'$  is the transpose of  $X$  and  $(X'X)^{-1}$  is the inverse of  $X'X$ . One can think of the observations on  $(X_1 \dots X_k, Y)$  as representing points in  $k+1$  dimensional space, to which one is fitting a  $k$ -dimensional hyperplane. The coefficient  $B_i$  on the variable  $X_i$  can be interpreted as the change in value of the dependent variable  $Y$  associated with a change of one unit in the variable  $X_i$  under conditions in which the other independent variables remain fixed in value. (The “other independent variables” that are assumed to remain constant

$$(7.1.7) \quad Y = XB' + U$$

where  $Y = \begin{vmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_n \end{vmatrix}$        $B = [B_0 \ B_1 \ B_k]$

$U = \begin{vmatrix} U_1 \\ \cdot \\ \cdot \\ U_n \end{vmatrix}$       and  $X = \begin{vmatrix} 1 & X_{11} & X_{12} & \cdot & X_{1k} \\ 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{n1} & \cdot & \cdot & X_{nk} \end{vmatrix}$

Figure 7.1.1

include  $U$ ; what this assumption amounts to in the case of  $U$  is discussed below.) For example, in the equation

$$(7.1.9) \quad Y = a_1X_1 + a_2X_2 + U$$

with  $Y$  measuring the height of individual plants in some population, and  $X_1$  and  $X_2$  the amount of water and fertilizer that these plants receive, the coefficient  $a_1$  will tell us what change in plant height is associated with a unit change in amount of water for a fixed level of fertilizer, assuming that the “other causes” of plant height represented by  $U$  also remain fixed.

## 7.2 Description, Prediction, and Causation

So far, I have described a set of techniques for fitting linear equations to a body of data. The question I now want to explore is what it means (or under what circumstances it is appropriate) to interpret such equations as describing causal connections between their independent and dependent variables.

In a recent paper, David Freedman (1997, p. 116) distinguishes three possible uses for an equation like (7.1.7)  $Y = B_1X_1 + \dots + B_nX_n + U$ :

- (i) To summarize or to describe or represent a body of data.
- (ii) To predict the value of a dependent variable  $Y$  from a set of independent variables  $X_1 \dots X_n$ .
- (iii) To predict the value of  $Y$  from an intervention that changes the value of  $X_1$ , etc.

Let us begin with (i). As we have seen, by construction, a regression equation like (7.1.7) describes the average or expected value of  $Y$  associated with the variables  $X_1 \dots X_n$  for a given body of data, assuming a linear functional form and the least squares criterion of best fit. In this sense, (7.1.7) will always represent in summary form a pattern or relationship in the data. Obviously, one can use (7.1.7) for this purpose even if one does not think that  $X_1 \dots X_n$  are causes of  $Y$ .

A second possible use for an equation like (7.1.7), as indicated in Freedman’s (ii), is to predict the value of  $Y$ , given various values of  $X_1 \dots X_n$  for other bodies of data besides the data on which (7.1.7) was estimated. Here again, successful prediction does not seem to require that the regression equation receive a causal interpretation. Instead, as Freedman argues (p. 116), what it required is that the process generating the data remain *stable* across time and space. To return to our example in which changes in atmospheric pressure  $A$  cause both drops in barometer readings  $B$  and the onset  $S$  of storms for some particular barometer/storm/pressure system, we may be able to use the value of  $S$  to successfully predict the value of  $B$  by regressing  $B$  on  $S$  (and omitting the causally relevant  $A$ ), even though there is no direct causal connection from  $S$  to  $B$ . As a less trivial example, consider that many biological and social

phenomena exhibit stable time trends of various sorts: in cold climates, trees lose their leaves in the fall and regain them in the spring, and in Western countries, retail sales tend to be highest in the fourth quarter because of religious holidays. A number of sophisticated statistical techniques falling under the general rubric of time series analysis make use of information about the past behavior of such trends to predict the future behavior of some variable of interest. Under the right conditions, such predictions can be extremely accurate. For example, an equation in which last year's fourth-quarter GNP figures as a predictor variable will often yield quite good predictions of this year's fourth-quarter GNP, but most of us will be reluctant to think of the former variable as a cause of the latter.

Freedman distinguishes very sharply between these first two uses of regression (to describe and to predict) and what he describes as (iii) causal inference. His view of what causal inference involves is essentially the view defended in previous chapters of this book. He writes, in a passage quoted in chapter 2:

Causal inference is different, because a change in the system is contemplated: for example, there will be an intervention. Descriptive statistics tell you about the correlations that happen to hold in the data; causal models claim to tell you what will happen to  $Y$  if you change  $X$ . Indeed, regression is often used to make counter-factual inferences about the past: what would  $Y$  have been if  $X$  had been different? (p. 116)

Elsewhere, he writes, in the context of a regression model like  $Y = BX + U$ , that we need to distinguish between two procedures:

Procedure #1. Select subjects with  $X = x$ ; look at the average of their  $Y$ 's.

Procedure #2. Intervene and set  $X = x$  for some subjects; look at the average of their  $Y$ 's.

These procedures are quite different. The first involves the data set as you find it. The second involves an intervention. (p. 118)

Freedman insists that it is this second procedure that is relevant to causal interpretation. We can think of his remarks as illustrating, in a regression context, the heuristic role of the manipulability conception of causation described in chapters 2 and 3. Freedman's claim is that if we are to interpret (7.2.1)  $Y = BX + U$  as describing a causal relationship running from  $X$  to  $Y$ , what we must mean is that if we were to *intervene* to change the value of  $X$ , then the value of  $Y$  would change in accord with the equation (7.2.1). Although Freedman does not go on to say explicitly what he means by an intervention, it should be clear that he must have in mind something like the characterization **IN** (chapter 3) if his procedure #2 is to be relevant to causal interpretation. By contrast, the use of a regression equation to describe or predict, as in (i) or (ii), carries with it no such implication about what would happen under an intervention. If we were to regress fourth-quarter GNP for a series of years (e.g., 1900–2002) on the previous year's GNP, we might obtain a relationships (call



it (G)) that could be used to accurately forecast fourth-quarter GNP for the coming year (2003). However, the fact that we could use (G) for successful prediction in this way provides us with no reason to believe that if we were somehow able to intervene to change 2002 fourth-quarter GNP, the relationship (G) would correctly describe how this intervention would change 2003 fourth-quarter GNP. As remarked above, the success of (G) in predicting fourth-quarter GNP depends on whether the mechanisms at work in the past in generating GNP remain stable or unchanged. But when we worry about the causes of GNP our concern is precisely, as Freedman says, with what would happen to present GNP if past GNP were not set by this generating mechanism but in a quite different way, via an intervention. Precisely because an intervention will break or change some previously existing causal mechanisms, we cannot automatically make the assumption of stability of the data-generating process that is required for extrapolation of past trends into the future.

Freedman's remarks illustrate a number of themes explored in earlier chapters. The contrast between procedure #1 when this is used predictively (e.g., one looks at past GNP and uses it to predict future GNP or at barometric pressure and uses it to predict storms) and procedure #2 is precisely the contrast between a backtracking counterfactual, the antecedent of which is not interpreted as an intervention, and a nonbacktracking counterfactual, the antecedent of which does correspond to an intervention. Alternatively, it is the distinction between conditioning and intervening, as these are characterized in chapter 2, or between the question of what I should believe about the value of  $Y$  given that I have observed that  $X = x$  and the question of what would happen to  $Y$  if I were to intervene to set  $X = x$ . As I have argued, it is the latter question that matters when it comes to causation and explanation.

Freedman's remarks also illustrate the fundamental difference between the question of whether a generalization is projectable (in the sense that it describes a correlation that, on the basis of limited data, we have good reason to expect will hold in the future or elsewhere) and the question of whether it describes a causal or invariant relationship. A generalization that is true for data so far observed will be projectable to other times and places if the mechanisms underlying the data are sufficiently stable and pervasive and the data themselves representative. Satisfaction of these conditions is not sufficient for the generalization to be invariant or describe a causal relationship. It is simply a mistake, although an extremely pervasive one in philosophy, to identify the question of whether a relationship is projectable with the question of whether it is causal or lawful.

I noted in chapter 1 that a number of philosophers have either denied that there is any interesting or principled contrast to be drawn between descriptive and explanatory claims or else have insisted that the business of science has to do purely with description and not with explanation. Regression (like the more general simultaneous equation techniques described below in this chapter) provides one important context in which these claims are clearly mistaken. Freedman's discussion draws our attention to a real and important difference between information of the sort described under (i) and (ii), which is descriptive

or predictive and carries with it no implications about what would happen if we were to intervene, and information described under (iii), which does have such implications and is causal or explanatory in character. At least in regression contexts, there is a clear distinction between using an equation to describe an association and using it to explain or to describe a causal relationship.

### 7.3 Regression and Causal Interpretation

I turn now to a closer look at some of the issues of causal interpretation surrounding (7.1.7)  $Y = B_1X_1 + \dots + B_nX_n + U$ . When (7.1.7) is interpreted as making a causal claim, it may be understood as claiming that each of  $X_1 \dots X_n$  are direct causes of  $Y$ —“direct” in the sense described in chapter 2. These causal relationships are understood as holding for each individual in the population of interest, in the sense that for each such individual, (7.1.7) characterizes the response of the value of  $Y$  possessed by that individual to some range of interventions that change the values of  $X_1 \dots X_n$  for that individual. On this interpretation, the error term  $U$  also has a causal interpretation: it represents the combined influence of all the other causes of  $Y$  besides  $X_1 \dots X_n$  that are not explicitly represented in (7.1.7). I call this the natural causal interpretation of (7.1.7). My interest in what follows is in using the ideas developed in previous chapters to specify in more detail what this natural causal interpretation involves, both in the case of individual equations like (7.1.7) and in the case of systems of equations. Just what should we understand (7.1.7) as claiming when we give it its natural causal interpretation?

I distinguish this interpretive issue from issues about the conditions under which one can reliably estimate the coefficients in (7.1.7). The latter issues are epistemological: they have to do not with what (7.1.7) means but with when and how one can determine the values of the coefficients in (7.1.7). Textbooks in econometrics have a great deal to say about such estimation problems. For example, as noted above, a familiar elementary result is that if the distribution of the error term satisfies conditions like (7.1.3a–d), then OLS estimators for the coefficients will have desirable properties such as unbiasedness. One of my themes is that a substantial amount of recent discussion, among both philosophers and econometricians, has run together interpretive and epistemological issues in connection with equations like (7.1.7). Conditions that have to do with reliable estimation, such as assumption (3d) regarding the uncorrelatedness of the error term with the independent variables in that equation, mistakenly have been taken as conditions that equations like (7.1.7) must satisfy if they are to have a causal interpretation.

The interpretive account I defend has a long history, although it often has been ignored or misunderstood in recent discussion. The basic themes can be found in econometricians like Frisch ([1938] 1995) and Haavelmo (1944), who emphasize the notions of invariance and autonomy and the role of hypothetical experiments in causal interpretation in interpreting structural equations. The core idea is also found in the remarks of David Freedman quoted

above. Among contemporary writers, Judea Pearl (e.g., 2000a) has articulated in a particularly clear and compelling way many of the ideas I will describe.

My contention, foreshadowed above, is that the causal claim represented by (7.1.7) is true if and only if, for some range of interventions on each of the variables  $X_i$  and  $U$ ,  $Y$  would change in just the way represented by (7.1.7) for each individual in the population. Thus, for each variable  $X_i$  there should be some possible intervention that changes the value of  $X_i$  by amount  $\Delta x_i^*$  and where the associated change in the value of  $Y$  is  $B_i \Delta x_i^*$ . (Recall that an intervention that changes  $X_i$  must not change or be correlated with changes in any of the other r.h.s variables in (7.1.7), including  $U$ .) To employ the terminology of chapter 6, the relationship (7.1.7) should be *invariant* under some interventions on  $X_i$  and  $U$ : it should remain stable under such changes. Here, invariance implies that the coefficients  $B_i$  remain constant and the linear functional relationship described by (7.1.7) is not altered by the intervention in question. It also implies that the values of the r.h.s. (independent) variables, including  $U$ , that figure in (7.1.7) do not change (or alternatively, that their *distribution* does not change; see below) under interventions on other r.h.s. variables. Thus, if, under an intervention that sets the variable  $X_i$  to some particular value  $x_i^*$ , the coefficient  $B_i$  or any other coefficient in (7.1.7) changes, or if the functional form of (7.1.7) changes, or if the value of  $X_j$  changes for  $j \neq i$ , then (7.1.7) is not invariant under this intervention and (7.1.7) does not correctly describe the causal relationship between  $X_i$  and  $Y$  for this value of  $X_i$ . (See below for a discussion of what the italicized phrase means.) If (7.1.7) is not invariant under any interventions on  $X_i$ , then it does not describe a causal relationship between  $X_i$  and  $Y$ . As we shall see, there are several varieties of invariance relevant to the causal interpretation of systems of equations. I call the notion just described *level invariance*, because it has to do with whether (7.1.7) is invariant under interventions that appropriately change the levels of its independent variables, including  $U$ .

Before proceeding, several clarifications are in order. First, on the interpretation put forward above, an equation like (7.1.7) makes a claim about the response of *each* individual in some population to interventions carried out on that individual: it is assumed that each such individual has the same response to interventions. However, assumptions about the distribution of  $U$  are clearly population-level assumptions that place a collective constraint on the values of  $U$  taken by all individuals in the population. The most natural way of bringing these two kinds of claims together is to think in terms of collective interventions that impose exactly the same intervention on some variable  $X_i$  for each individual in the population; claims about the invariance of  $U$  under interventions on  $X_i$  can then be understood as claims about what the individual values of  $U$  would be (called *value invariance* below) or alternatively, claims about what the distribution of  $U$  would be (*distribution invariance*) under such a collective intervention.<sup>3</sup>

Second, I emphasize that the above remarks are put forward (only) as an account of what it is for a particular kind of equation or mathematical

relationship—namely, a regression equation of the form (7.1.7)—to correctly represent a quantitative causal relationship. They are *not* put forward as a completely general account of what it is for a relationship to be causal. Obviously, it can be true that  $X_i$  causes  $Y$  even if (7.1.7) is not invariant under interventions on  $X_i$ . This will happen if the relationship between  $X_i$  and  $Y$  is causal but does not conform, in its quantitative behavior, to (7.1.7), for example, if the relationship is nonlinear. However, it would be a mistake to conclude from this that the account offered above fails to capture what must be true for an equation like (7.1.7) to describe a causal relationship. Although (7.1.7) says, when interpreted causally, that each of the independent variables  $X_i$  causes  $Y$ , it does not *just* say this. Rather, it makes a much more specific quantitative claim about the causal relationship between  $X_i$  and  $Y$ , and it is this quantitative claim that we are trying to capture.

Third, my aim is to describe conditions that must be satisfied if an equation like (7.1.7) is to be true under its natural causal interpretation. Whether we have good reason to believe the causal claim made by this or that particular regression equation is a separate question that needs to be settled on a case-by-case basis. It is no part of my view that regression equations in the social and biomedical sciences are usually or often correct when given the causal interpretation described above.

A fourth point, which should be familiar from chapter 6, is that even if a relationship such as (7.1.7) is invariant for some range of values of its independent variables that are set by interventions, it is unlikely to be so for others. Consider again (7.1.9)  $Y = a_1X_1 + a_2X_2 + U$  with  $Y$  interpreted as plant height and  $X_1$  and  $X_2$  the amount of water and fertilizer an individual plant receives. Even if it is true that (7.1.9) is invariant under some range of interventions on the amount of water a plant receives, it is clearly not invariant under all such interventions: one cannot make a plant grow arbitrarily tall by putting arbitrarily large amounts of water on it. Following the argument of chapter 6, we may accommodate this observation by explicitly relativizing the notion of invariance (and, correlatively, the connection between causation and successful prediction of the outcome of a hypothetical experiment) to a range of values of variables set by interventions. In other words, we think of (7.1.9) as invariant under some interventions but not others and as correctly describing the results of some hypothetical experiments but not others. As before, I assume that to qualify as a correct causal description, a relationship must be invariant under some range of interventions but need not be (and, in the case of the sorts of causal relationships typically described by regression equations, will not be) invariant under all interventions. Thus, for example, (7.1.9) may qualify as a true causal generalization (with respect to those values of its independent variables for which it is invariant) if it is invariant under interventions that set the value of the amount of water that each plant in some population receives to 1, 2, or 3 liters, even if (7.1.9) would break down if the plants were to receive 1,000 liters of water. In such circumstances, (7.1.9) correctly describes how, by manipulating the amount of water a plant receives within the 1–3-liter range, one may manipulate its height, and this is enough

for (7.1.9) to be a true causal description within this range. I suggest below that when they are invariant at all, the range of interventions over which structural equations in the social sciences are invariant is often rather narrow and may in fact be confined to the population on which the equation is estimated or closely similar populations.

Fifth, my argument has been that a necessary and sufficient condition for an equation like (7.1.9) to correctly represent a causal relationship is that it be invariant under some range of interventions on its r.h.s. variables. However, again following chapter 6, I do *not* claim that when (7.1.9) represents a causal relationship it will be invariant *only* under interventions on such variables. Typically, when (7.1.9) represents a causal relationship, it will be stable under many other sorts of changes as well, including changes in background conditions (factors not included in (7.1.9), either among the measured variables or in the error term) and changes in variables that explicitly occur in (7.1.9) where those changes do not involve interventions.

In connection with the second possibility, consider some process that alters both  $X_1$  and  $X_2$  in (7.1.9). Such a process will not count as an intervention on  $X_1$ , because the change produced in  $X_1$  will be correlated with the change produced in another cause of  $Y$ , namely  $X_2$ , and by parallel reasoning also will not count as an intervention on  $X_2$ . Nonetheless, we expect that if (7.1.9) describes a causal relationship, it will be stable or invariant under some range of such changes. It is important to understand what this means: in saying that (7.1.9) will be invariant under such changes, I mean that the change in  $Y$  will be what (7.1.9) says it will be, given the changes in  $X_1$  and  $X_2$ . Thus, if  $X_1$  is changed by amount  $\Delta X_1$  and  $X_2$  by amount  $\Delta X_2$ , the total change in  $Y$  will be  $a_1\Delta X_1 + a_2\Delta X_2$  if (7.1.9) is invariant under this change. This contrasts with the change in  $Y$  that would be produced by an intervention on  $X_1$  which is just  $a_1\Delta X_1$ .

To put the point slightly differently, we may interpret the individual coefficients in a linear regression equation as telling us what change in  $Y$  would be produced by interventions on the associated r.h.s. variables; thus, the coefficient  $a_i$  tells us what the change in  $Y$  would be under an intervention that produces a unit change in  $X_i$ . When a process occurs that changes several of the r.h.s. variables in (7.1.9) simultaneously, (7.1.9) will often remain invariant, but in such cases, the total change in  $Y$  will be the net effect or sum of the contributions made by the changes in each of the independent variables where each of these contributions is the change in  $Y$  that would have occurred if an intervention had occurred on just that variable. Thus, what is “the same” across cases in which  $X_1$  is altered by amount  $\Delta X_1$  by an intervention and cases in which  $X_1$  is altered by amount  $\Delta X_1$  and  $X_2$  is altered by amount  $\Delta X_2$  is the *causal contribution* in the sense described in chapter 2 (namely,  $a_1\Delta X_1$ ) made to the total change in  $Y$  by the change in  $X_1$  or, alternatively, the relationship (7.1.9). Both of these need to be distinguished from the total change in  $Y$ , which will not, of course, be the same in these two cases. (Recall the discussion of direct, contributing, and total causes in chapter 2 and the need to be clear about just which relationship is invariant, emphasized in section 6.3.)

I can further clarify the connection between causation and invariance I advocate by contrasting my views briefly with those of Nancy Cartwright. In recent work, Cartwright (e.g., 1995) suggests that we should distinguish between two questions: whether a relationship is causal, and the extent to which it is stable or invariant across various sorts of changes. She contrasts what she calls “mere causal relationships” with “capacities that will be stable across a range of envisioned changes” or, as she also describes them, “*super-causal relations*” which “remain invariant across a range of envisioned interventions” (p. 55). A relationship can be causal without being “supercausal” or “invariant” (p. 56). Situations meeting the conditions for a controlled experiment will sometimes allow us to establish “what causal relationships obtain in that situation,” but because causation is different from invariance we cannot infer from this “what causal relationships will obtain outside that situation.” Additional assumptions about capacities or invariance are required to “export” causal conclusions to different situations.

Where Cartwright sees a sharp distinction between two questions (Is this relationship causal? Is it invariant?), I see these questions as deeply intertwined. As argued above, if one accepts the view that causal relationships tell us something about the results of hypothetical experiments, then it follows that some measure of invariance under interventions is necessary for a relationship to be causal at all. From this perspective, what relatively stable or invariant (or supercausal) relationships have is not some additional feature that is not present at all in “mere” causal relationships, but more of the same feature: invariance and stability under a larger or more important range of interventions and changes than is present in mere causal relationships.

Let us now turn to a more detailed look at the role of the error term in equations like (7.1.7). To begin with, it is important to understand that if (7.1.7) is to be given a causal interpretation, we must think of  $U$  as representing the combined upshot of the *causes* of the dependent variable  $Y$  that are uncorrelated with the r.h.s. variables in (7.1.7). Suppose instead that we adopt what might be called a purely *correlational* interpretation of the error term: we think of the error term as just representing some set of factors (not necessarily causes of  $Y$ ) that obey the assumptions (7.1.3a–d). On such an interpretation, we may make the error term automatically satisfy the uncorrelatedness assumption by construction; we need only define it as the residual  $u_i = y_i - (B_0 + B_1x_{1i} + B_2x_{2i} + \dots + B_kx_{ki})$  or, more compactly, as  $U = Y - BX$ , where the coefficients in (7.1.7) are defined by their OLS estimates (cf. Woodward 1999). Clearly, if the error term is identified with the residual in this way, any assumptions about its distribution will no longer be a substantive constraint. It is only if the error term is understood as standing for something real and independent (omitted causes) that there can be serious empirical issues about whether (7.1.3a–d) are satisfied and about the invariance of the error term. I assume such a substantive, causal interpretation of the error term in what follows.<sup>4</sup>

I remarked above that when we claim that a regression equation is level-invariant, this implies that some appropriate invariance condition must hold

for the error term  $U$  as well as for the relationship between the variables  $X_i$  and  $Y$ . In an illuminating recent paper, David Freedman (forthcoming) distinguishes two different kinds of invariance claims one might be prepared to make about  $U$ . One kind of claim has to do with invariance under changes in the *individual values* assumed by  $U$  on particular occasions. That is, it is assumed that (7.1.7) would continue to hold under (at least some range of) changes in the value of  $U$  and also that interventions that change the value of each of the r.h.s variables in (7.1.7) will not change the value of  $U$ , and conversely. As Freedman remarks, invariance assumptions of this sort will license counterfactual inferences about what the value of  $U$  would have been if an intervention (within the range of invariance of (7.1.7)) changes the value of  $X_i$  on some particular occasion (under such an intervention the value of  $U$  would not change) and hence also inferences about what the change in  $Y$  would have been under this intervention. Let us call these *value invariance assumptions* about  $U$ .

An alternative, more modest set of invariance assumptions has to do instead only with the *distribution* of  $U$ : it is assumed that (7.1.7) would continue to hold under changes in the distribution of  $U$ , that interventions that change the distribution of  $U$  will not change the values of the other r.h.s. variables in (7.1.7), and that interventions on those variables will not change the distribution of  $U$ . For example, if the distribution of  $U$  is normal and  $U$  is uncorrelated with any of the r.h.s. variables in (7.1.7), it is assumed that (7.1.7) would continue to hold if the distribution of  $U$  changes so that it is no longer normal or in such a way that  $U$  is now correlated with  $X_i$ . Similarly, an intervention on  $X_i$  will not change the distribution of  $U$  from normal to nonnormal or change a distribution from one in which  $U$  is not correlated with  $X_i$  to one in which it is correlated. Let us call the assumptions just described *distributional invariance assumptions*. Such distributional invariance assumptions about  $U$  allow for inferences about how the probability that  $Y$  will assume some value would change under interventions that change the values of the  $X_i$  but not inferences as to how the precise value of  $Y$  would change under such interventions.

Presumably, the value invariance assumptions concerning  $U$  imply the distributional invariance assumptions about  $U$ , but not conversely. Provided that we take the deterministic character of (7.1.7) literally, value invariance seems very natural. Indeed, I tacitly assumed it above when I claimed that the coefficients in a regression equation tell us how the value of  $Y$  and not just the probability distribution of  $Y$  would change under interventions on  $X$ .<sup>5</sup> However, for the claims that follow about the significance of the uncorrelatedness of the error term, it is only necessary to assume that, for an equation like (7.1.7) to have a causal interpretation, the distributional invariance assumptions concerning  $U$  must be satisfied.

If we agree that, with respect to the error term, the key to causal interpretability is the satisfaction of (at least) distributional invariance, an important consequence follows. Contrary to what many writers (including both philosophers and statisticians<sup>6</sup>) have maintained, it is *not* necessary, if a

regression equation is to have a causal interpretation or to accurately describe a set of causal relationships, that the error term in the equation be uncorrelated with each of the independent variables in that equation. What matters is not the actual distribution of  $U$  (and, in particular, whether  $U$  satisfies the uncorrelatedness assumption), but rather the answers to such questions as whether the equation is invariant (*would* continue to hold) under *changes* in the distribution of  $U$  or whether interventions on some of the other r.h.s. variables in the equation would *change* the distribution of  $U$ . Suppose that we are estimating a regression equation of the form (7.1.7) for a certain population and that the uncorrelatedness assumption is satisfied. The idea that if this equation correctly represents a causal relationship it should be invariant under changes in the distribution of the error term implies, among other things, that if the distribution of the omitted causes  $U$  of  $Y$  changes in such a way that the error term now is correlated with one or more of the independent variables, the relationship (7.1.7) between  $Y$  and  $X_1 \dots X_n$ ,  $U$  should nonetheless continue to hold (in the sense that this relationship should continue to describe what would happen to  $Y$  under interventions on  $X_1 \dots X_n$ ,  $U$ ). Because the error term is now correlated, one will no longer be able to use OLS estimation techniques (although one may still be able to use other estimation techniques), but assuming that (7.1.7) is invariant under this change, it will be just as much a correct causal representation as before. In other words, the role of the uncorrelatedness assumption is purely epistemological: it is a necessary condition for certain estimators for the coefficients in (7.1.7) to be unbiased but not a necessary condition for causal interpretability.

## 7.4 Systems of Equations and Modularity

Regression equations represent a particularly simple causal structure in which a single dependent variable is causally influenced by one or more independent variables but no causal relationships are represented among the independent variables themselves and no reciprocal or cyclic causal links back from the dependent variable onto the independent variables are represented. Often, however, social scientists and other users of causal modeling techniques want to represent more complex structures. This is accomplished through the use of systems of equations. The conventions for interpreting such equations causally parallel those for single regression equations. Each equation in a system contains a single dependent variable on the l.h.s. and one or more r.h.s. variables which are interpreted as the direct causes of the l.h.s. variable. If one wishes to represent causal relationships among the r.h.s. variables, one adds additional equations conforming to the convention just described. For example, in the system of equations

$$(7.4.1) \quad Y = aX + U$$

$$(7.4.2) \quad Z = bX + V$$



(7.4.1) says that  $X$  is a direct cause of  $Y$  and (7.4.2) says that  $X$  is a direct cause of  $Z$  (as before,  $U$  and  $V$  are error terms that represent causes of the dependent variable in each equation that are unmeasured and not explicitly represented). As we did in chapters 2 and 3, we may also represent the structure (7.4.1)–(7.4.2) by means of a directed graph, following the convention that an arrow directed out of one vertex and into another means that the former is a direct cause of the latter (see figure 7.4.1).

In view of our earlier discussion, it is natural to impose the requirement that if a system of equations like (7.4.1)–(7.4.2) correctly represents the causal facts, then each individual equation (7.4.1) and (7.4.2) must be level-invariant under some range of interventions on the r.h.s. variables of that equation. This means, of course, that (7.4.1) will correctly describe how  $Y$  will change under some interventions on  $X$  and (7.4.2) will describe how  $Z$  will change for some interventions on  $X$ .

Suppose, however, that although both (7.4.1) and (7.4.2) are level-invariant, all possible ways of changing  $Y$  (including interventions on  $Y$ ) disrupt (7.4.2) in the sense of changing the relationship between  $X$  and  $Z$  described by (7.4.2). This suggests a second invariance requirement, distinct from the requirement of level invariance, which it is natural to impose on systems of equations like (7.4.1)–(7.4.2). This is the requirement of *modularity*, introduced briefly in chapter 2. The reader may recall the basic idea: ideally, each equation in a system of equations should represent a distinct causal mechanism, where the criterion for distinctness of mechanisms is that distinct mechanisms should be changeable (in principle) independently of one another. This in turn means that it should be possible to alter or disrupt each of the equations in the system (i.e., by altering the operation of the mechanism with which it is associated) without altering or disrupting the others. Applied to (7.4.1)–(7.4.2), this means that if these equations correctly describe the causal relationships in the system they represent, then the mechanism or causal relationship by which  $X$  affects  $Y$  should be distinct from the mechanism by which  $X$  affects  $Z$ , and this in turn means that each mechanism should be disruptable independently of the other. A system of equations having this feature will be modular.<sup>7</sup>

To be more specific, consider a process that fixes  $Y$  to some value  $y$  in such a way that  $Y$  is no longer influenced by  $X$  and  $U$ . We can think of this as replacing (7.4.1) with a new equation (7.4.1\*)  $Y = y$  representing the fact that the value of  $Y$  is now fixed at  $y$  rather than, as was previously the case, being determined by  $X$  and  $U$ . If, under all such changes in  $Y$ , equation (7.4.2) is

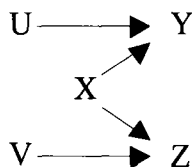


Figure 7.4.1

disrupted, then the system (7.4.1)–(7.4.2) will not be modular. Moreover, if (7.4.2) is disrupted whenever (7.4.1) is replaced by an equation of the form (7.4.1\*), then the value of  $Z$  will change under these changes in  $Y$  even when the value of  $X$  does not change and even though (7.4.1)–(7.4.2) claims there is no causal connection between  $Y$  and  $Z$ . This in turn suggests that the representation (7.4.1)–(7.4.2) is inadequate or incomplete in some way; for example, perhaps there is a causal relationship connecting  $Y$  to  $Z$  that is not represented by (7.4.1)–(7.4.2). More generally, we can say that if all changes that alter one equation also alter some other equations in a system, then the system will be misspecified in the sense that it will fail to correctly and completely represent the causal structure that it purports to model: variables will change in response to changes in other variables even though the equations represent the variables as causally unrelated or, alternatively, will fail to change under changes in the values of other variables in the way that the equations suggest that they should.

These considerations can be used to motivate the following proposal: when a system of equations correctly represents causal structure, not only should each equation in the system be level-invariant, but for each equation there should be some possible change that alters that equation or replaces it with another equation while leaving the other equations in the system unaffected. We may think of changing an equation (or the mechanism(s) or relationship(s) represented by it) as a matter of intervening on the l.h.s. (dependent) variable in the equation so that the value of that variable is now fixed by the intervention rather than by whatever variables previously determined its value. When I say that the other equations are unaffected, I mean that they would continue to hold and continue to be level-invariant under this change. We thus have:

**MODULARITY.** A system of equations is *modular* if (i) each equation is level-invariant under some range of interventions and (ii) for each equation there is a possible intervention on the dependent variable that changes only that equation while the other equations in the system remain unchanged and level-invariant.

I have defined modularity in such a way that if a system is modular, then each equation in the system must be level-invariant. Nonetheless, modularity and level invariance are distinct concepts. In particular, it is not true that if each equation in a system of equations is level-invariant the system must be modular. (As we shall see below, even if the so-called reduced-form equations associated with a system of equations are level-invariant, they need not be they not be modular.) Intuitively, level invariance is a condition that applies *within* individual equations and concerns whether an individual equation is invariant under interventions on its r.h.s. variables. By contrast, modularity is an invariance condition that also applies *between* equations and has to do with whether each equation is invariant under changes in other equations: it is an “equation invariance” condition. The distinction between level invariance and equation invariance provides a concrete illustration of the more general

point, urged in chapter 6, that there are a variety of different invariance conditions, corresponding to the stability of different relationships under different sorts of changes. These conditions are related to the truth of different counterfactuals and illustrate the point that a variety of counterfactuals rather than just a single counterfactual schema are relevant to questions of causal interpretation. Whereas level invariance has to do with the truth of counterfactuals describing how the dependent variable in some equation would change under interventions on the independent variables in that equation, modularity has to do with (different) counterfactuals concerning whether the relationship described by one equation would continue to hold as other equations are disrupted.

We can bring out more clearly what modularity involves by considering the following system of equations:

$$(7.4.3) \quad Y = aX + U$$

$$(7.4.4) \quad Z = bX + cY + V$$

Let us now rewrite (7.4.3) and (7.4.4) as follows:

$$(7.4.3) \quad Y = aX + U$$

$$(7.4.5) \quad Z = dX + W$$

where  $d = b + ac$  and  $W = cU + V$ .

Because (7.4.5) is obtained by substituting (7.4.3) into (7.4.4), the system (7.4.3)–(7.4.4) has exactly the same solutions in  $X$ ,  $Y$ , and  $Z$  as the system (7.4.3)–(7.4.5). Because  $X$ ,  $Y$ , and  $Z$  are the only measured variables, (7.4.3)–(7.4.4) and (7.4.3)–(7.4.5) are in a sense (to be discussed in more detail below) “observationally equivalent”: they imply or represent exactly the same facts about the patterns of correlations that obtain so far among these measured variables. Nonetheless, by the rules given above for interpreting systems of equations, these two systems correspond to different causal structures. (7.4.3)–(7.4.4) says that  $X$  is a direct cause of  $Y$  and that  $X$  and  $Y$  are direct causes of  $Z$ . By contrast, (7.4.3)–(7.4.5) says that  $X$  is a direct cause of  $Y$  and that  $X$  is a direct cause of  $Z$  but says nothing about a causal relation between  $Y$  and  $Z$ . This difference is also reflected in the graphical representation associated with the two systems (figures 7.4.2 and 7.4.3).

Despite their observational equivalence, if (7.4.3)–(7.4.4) is modular, then (7.4.3)–(7.4.5) cannot be (and vice versa). To see this, consider an intervention on the variable  $Y$  in (7.4.3) that replaces (7.4.3) with the new equation (7.4.3\*)  $Y = y$ . In effect, what this intervention does is to set the coefficient  $a$  in (7.4.3) equal to 0. If the system (7.4.3)–(7.4.4) is modular, (7.4.4) will continue to hold under at least one such change in (7.4.3). By contrast, if (7.4.3)–(7.4.4) is modular, (7.4.5) must change under this intervention because, as we

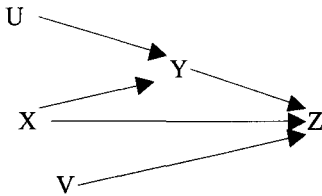


Figure 7.4.2

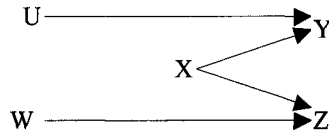


Figure 7.4.3

have seen, its effect is to change the value of the coefficient  $a$  in (7.4.3) and the coefficient  $d$  in (7.4.5) is a function of  $a$ . Thus, changing  $a$  in (7.4.3) will change  $d$  and hence (7.4.5). This corresponds to our judgment that if (7.4.3)–(7.4.4) is a correct representation of the causal facts, then (7.4.3)–(7.4.5) collapses or mixes together distinct mechanisms or causal routes—the influence of  $X$  on  $Z$  that occurs because  $X$  directly influences  $Z$  (this is represented by the coefficient  $b$ ) and the influence that occurs because  $X$  influences  $Y$ , which in turn influences  $Z$  (this is represented by the product  $ac$ )—into a single overall mechanism linking  $X$  and  $Z$ , which is represented by the coefficient  $d$ . This failure to correctly segregate the system being modeled into distinct mechanisms is directly reflected in the nonmodularity of (7.4.3)–(7.4.5).

We can also bring out the difference between (7.4.3)–(7.4.4) and (7.4.3)–(7.4.5) in a slightly different way. If (7.4.3)–(7.4.4) is a correct representation of the causal facts (and hence is modular), we know that there is an intervention under which (7.4.3) will be disrupted but (7.4.4) will not be. Suppose that this intervention consists of setting  $Y$  to some new value  $k$ . Under this intervention,  $Y$  will continue to make just the contribution to  $Z$  that is indicated by (7.4.4); that is, the contribution will be  $ck$ . Thus, according to (7.4.3)–(7.4.4), this intervention on  $Y$  will change the value of  $Z$ . By contrast, according to (7.4.3)–(7.4.5),  $Y$  does not cause  $Z$  and hence there are no interventions on  $Y$  that will change  $Z$ . (Recall that an intervention on  $Y$  should be uncorrelated with other causes of  $Y$  such as  $X$ .) Assuming that (7.4.3)–(7.4.4) is the correct structure, what will happen under this intervention only, is (as we have seen) that the coefficient  $d$  will change in (7.4.5) so as to reflect the change in the value of  $Z$  that is really produced by the change in the value of  $Y$ , but this fact about the dependence of  $Z$  on  $Y$  is not represented in the equations (7.4.3)–(7.4.5). Thus, assuming that (7.4.3)–(7.4.4) is modular, (7.4.3)–(7.4.5) fails to correctly represent what will happen under hypothetical interventions on  $Y$ . Note that this is not a failure of level invariance in the sense that (7.4.3)–(7.4.5) represent a relationship as causal that fails to be level-invariant. Instead, the problem is that there is a causal or level-invariant relationship that (7.4.3)–(7.4.5) fail to represent.

There are many other systems of equations besides (7.4.3)–(7.4.5) that may be obtained from (7.4.3)–(7.4.4) by equality-preserving algebraic transformations

and are in this sense observationally equivalent to the latter. For example, we can also rewrite (7.4.3)–(7.4.4) as

$$(7.4.3) \quad Y = aX + U$$

$$(7.4.6) \quad Z = eY + R \text{ where } e = (b/a + c) \text{ and } R = V - (b/a)U$$

Again, if (7.4.3)–(7.4.4) is modular, (7.4.3)–(7.4.6) will not be, because changing  $a$  will change the value of  $e$ . Again, (7.4.3)–(7.4.6) represents a different set of causal claims from (7.4.3)–(7.4.4) because it makes different predictions about what will happen under various hypothetical changes. If one accepts that, despite their observational equivalence, at most one of these systems can correctly represent the causal facts, there must be some additional constraint that is satisfied by the correct system. Modularity is the natural candidate for this constraint. The idea is that among all of the observationally equivalent representations, we should prefer the one that is modular because it will be the one that correctly and fully represents causal relationships and mechanisms. As Alderich (1989) puts it, the constraint of modularity (or, as he calls it, “autonomy”) picks out a “privileged parameterization.”

The equations (7.4.3)–(7.4.5) are the *reduced-form equations* associated with the system (7.4.3)–(7.4.4). In general, one forms the reduced-form equations associated with a system by first identifying the exogenous variables in the system, that is, the variables that are not themselves caused by any of the other variables in the system and do not have arrows directed into them in the graphical representation of the system. One then substitutes into the equations in the system in such a way that one is left with a set of equations, one for each endogenous variable, which have the endogenous variable on their l.h.s. and only exogenous variables (and an error term) on their r.h.s. It is always possible to do this, and the resulting reduced-form system will always be observationally equivalent to the original system regardless of its structure. Moreover, the error term in each equation will always be uncorrelated with the r.h.s. variables in that equation, and hence one may always estimate the values of the coefficients in the reduced-form equations by OLS. By contrast, it is not true for all systems of equations that the values of the coefficients are estimable from statistical data about the measured variables; in the jargon of econometrics, some of the coefficients may be *unidentifiable*. In addition to this, as long as no changes occur in the coefficients in the original system (either because of interventions on the endogenous variables or for some other reason), the reduced-form equations will be level-invariant (under interventions on the exogenous variables in each equation) if the original system is modular. If we care only about producing an observationally adequate representation of the pattern of correlations among the measured variables  $X$ ,  $Y$ , and  $Z$ , or if we care only about finding an observationally adequate representation that is level-invariant in the exogenous variable  $X$ , we may use just the reduced-form equations.

However, as the preceding discussion illustrates, if the original system is modular, the associated reduced-form equations will not be modular. For example,

if (7.4.3)–(7.4.4) are modular (7.4.3)–(7.4.5) will not be. We thus see that, as claimed above, it is possible to have a system of equations all of which are level-invariant but that fails to be modular. Again, if researchers are not always content with the reduced-form representation—and it is clear that they are not—this can only be because they value something else (modularity) besides level invariance and observational adequacy.

I can further bring out the significance of this last point by connecting it with the passage from Haavelmo's monograph (1944), which is quoted in chapter 6. It will be recalled that Haavelmo envisions a researcher who investigates the relationship ( $R$ ) between the maximum speed attained by a particular make of car and the depression of the gas pedal as this "experiment" is repeated under exactly the same conditions: the same road conditions, fuel mixture, state of the engine, and so on. It is plausible that, provided these conditions continue to obtain, ( $R$ ) will be level-invariant under some range of interventions that depress the gas pedal. Because of this—because ( $R$ ) describes how, within this range, one can use the pedal to manipulate the speed—( $R$ ) qualifies as a genuine causal relationship. Nonetheless, as Haavelmo says, ( $R$ ) strikes us as explanatorily shallow and scientifically uninteresting. Haavelmo contrasts ( $R$ ) in this respect with an engineering style theory, we called it ( $E$ ), in which the operation of the car is decomposed into a number of distinct mechanisms and the principles governing these. We expect that ( $E$ ) will be modular in the sense that it will consist of independently changeable representations of the operations of mechanisms that are in fact distinct from one another. Thus, we expect that we can change or interfere with the operation of one mechanism making up the engine (e.g., the spark plugs) without automatically changing the operation of other mechanisms (e.g., the relationship between the air pressure in the tires and friction with the road surface). It is this feature that allows us to trace out the implications of hypothetical changes in the operation of the various components of the car, just as the modular representation (7.4.3)–(7.4.4) allows us to trace out the implications of changes in the individual mechanisms it represents. A theory like ( $E$ ) should thus enable one to see how, if the operation of any of the mechanisms making up the car were to change (e.g., the spark plugs are cleaned or more air is put in the tires) or if various factors in the environment (the grade of the road, the head wind, etc.) were to change, the operation of the car and the relationship between the gas pedal and the speed would change. Whereas each of the individual equations in ( $E$ ) will be invariant under changes in the other equations, ( $R$ ) will fail to be invariant under most such changes. We may recall that Haavelmo says that because ( $R$ ) is invariant under a smaller set of changes and interventions than the equations in ( $E$ ), it is less *autonomous* than those equations, and he links this to the fact that ( $R$ ) is less satisfactory from the point of view of explanation and causal understanding than  $E$ . An autonomous (or relatively autonomous) system of equations will be a system that is modular (or relatively modular).

I take Haavelmo to be suggesting that we should think of ( $R$ ) as like a non-modular reduced-form equation. Just as the coefficient  $d$  in equation (7.4.5)

(on the assumption that (7.4.3)–(7.4.4) is modular and (7.4.3)–(7.4.5) is not), represents a sum or mixture of the coefficients corresponding to several distinct mechanisms, so (R) summarizes an overall relationship between speed and gas pedal position that is the combined upshot of the operation of many different mechanisms making up the car. Like (R), (7.4.5) will be invariant under some interventions that change the level of its exogenous variable  $X_1$  and hence will qualify as causal. But both (R) and (7.4.5) will continue to hold only as long as none of the many causal relationships or mechanisms that contribute to the overall relationship they describe are altered. For example, a change in any one of the coefficients  $b$ ,  $a$ , or  $c$  will change the relationship described by (7.4.5), just as a change in the any one of the mechanisms making up the car engine will disrupt (R). Just as (R) is less autonomous (less invariant) than (E), so (7.4.5) is thus less autonomous than, say, (7.4.4) (again on the assumption that (7.4.3)–(7.4.4) is the modular representation) in the sense that interventions that disrupt (7.4.4) will also disrupt (7.4.5) but there are interventions—for example, interventions that disrupt (7.4.3)—that will disrupt (7.4.5) but not (7.4.4). Just as we believe that (E) is more satisfactory than (R) from the point of causal explanation, so we should prefer (7.4.3)–(7.4.4) to (7.4.3)–(7.4.5).

Before proceeding, let me address some possible misunderstandings of the argument of the preceding paragraphs. First, there are, of course, many causal systems for which there are at present no technologically feasible methods that allow for separate interference with all of the distinct mechanisms that compose them. What was said in previous chapters about interventions and invariance applies here as well: when we ask whether there is an intervention that disrupts one equation that would leave other equations unchanged, what we are interested in is what would happen under certain hypothetical possibilities and not whether such interventions are practically possible. In the example discussed above, what makes (7.4.3)–(7.4.5) non-modular (if (7.4.3)–(7.4.4) is modular) are the mathematical relationships between the coefficients of the two systems. It is these that ensure that if the coefficient  $a$  in (7.4.3) changes, the coefficient in (7.4.5) must change, that is, that there is not even a hypothetical intervention that changes (7.4.3) without changing (7.4.5).

Second, I emphasize that the argument is conditional in character: if (7.4.3)–(7.4.4) is modular, then (7.4.3)–(7.4.5) will not be modular and will fail to correctly represent the results of various hypothetical interventions. Because the coefficients  $a$ ,  $b$ , and  $c$  in (7.4.3)–(7.4.4) can also be written as functions of  $a$  and  $d$  in (7.4.3)–(7.4.5), a parallel argument could also be used to show that if (7.4.3)–(7.4.5) is modular, then (7.4.3)–(7.4.4) will not be. What the argument shows is that, at most, one among the various alternative systems of equations relating  $X$ ,  $Y$ , and  $Z$  that are observationally equivalent can be modular and hence that modularity represents a real constraint on the choice of a system of equations. However, the argument does not purport to tell us which, if any, of these alternative systems is in fact the modular one. It is nature or the world—and, in particular, facts about what the causal mechanisms are and about what would happen under different hypothetical changes—that

determines this. In other words, researchers must first determine, in some independent way, which mechanisms are distinct and what would happen under various hypothetical interventions. They may then represent this information by a system of equations, guided by conventions of the sort described above: that different equations should correspond to distinct mechanisms, that if an intervention on  $X$  would change  $Y$ , then  $X$  should occur on the r.h.s. of an equation and  $Y$  on the l.h.s., and so on.

These observations may also help to address another concern that may trouble some readers: (7.4.3)–(7.4.5) has been obtained from (7.4.3)–(7.4.4) by a series of equality-preserving algebraic transformations. In view of this, how can it be true that these two systems of equations represent different causal claims? Aren't the two systems "mathematically equivalent"? If (7.4.3)–(7.4.4) and (7.4.3)–(7.4.5) said only that the quantities  $X$ ,  $Y$ , and  $Z$  are regularly associated or correlated in a certain pattern, then they would indeed be interchangeable representations of the same set of correlational facts. However, when interpreted causally, systems like (7.5.3)–(7.5.4) say more than this. As Judea Pearl (2000a), among others, has emphasized, the syntactic form of these equations also conveys information: information about what would happen under hypothetical interventions. For example, when we write  $Y$  on the l.h.s. of an equation and  $X$  on the r.h.s., we convey the information that an intervention on  $X$  would change  $Y$  but not the information that an intervention on  $Y$  will change  $X$ . This is so even though we can derive (7.4.6)  $X = (1/b)Z - (c/b)Y$  from (7.4.4)  $Z = bX + cY$  and vice versa. Similarly, writing (7.4.3)–(7.4.4) rather than (7.4.3)–(7.4.5) conveys, through the syntactic convention that different equations should represent distinct mechanisms, one set of claims about what the causal mechanisms are rather than another and an accompanying set of claims about what will happen under hypothetical interventions. It is because such syntactic information is lost or changed under algebraic manipulation that it is possible for these two systems to represent different sets of causal facts. The fact that algebraic transformation of a correct representation of a set of causal relationships can lead to an incorrect representation is yet another illustration of the more general point, urged in previous chapters, that truth-preserving mathematical transformations and derivations need not automatically mirror causal relationships.

I said above that (7.4.3)–(7.4.4) and (7.4.3)–(7.4.5) were, in a sense, observationally equivalent. The preceding paragraphs enable us to be somewhat more precise about what this means and to understand how, despite this fact, they can represent different causal structures. Assuming that their coefficients and error terms are related in the way described, the two systems agree about the actual patterns of correlations among the measured variables that obtain so far. What they disagree about is modal or counterfactual: they make inconsistent predictions about what would happen should various changes or interventions occur. It is entirely possible, of course, that these changes will not occur, in which case the two systems will continue to agree about what will be observed. When we take the two systems to disagree about the causal facts, we accept that there is a fact of the matter about what would happen



under various hypothetical changes even if these never occur and that it matters which of the two sets of equations makes correct predictions about this.

We may contrast this point of view with the idea, associated with van Fraassen's (1980) constructive empiricism, that the aim of scientific theorizing, or at least the standard by which scientific theories should be judged, has to do only with the accurate representation of what actually happens in the realm of observables. Constructive empiricism has been widely criticized on the grounds that the observable/nonobservable distinction is both problematic and lacks the epistemic significance assigned to it by van Fraassen. From the point of view of our present discussion, however, there is an additional, distinct problem: the commitment of constructive empiricism to the idea that scientific theories need be adequate only to what actually happens (whether this involves observable or nonobservable happenings) is also problematic. The contrary view for which I have been arguing is that a legitimate aim of science (and an aim that scientific theories *should* have insofar as they aim at the representation of causal relationships) is the accurate representation of what would happen under various counterfactual eventualities and not just what actually happens. The structural equations literature embodies this point of view.<sup>8</sup>

The ideas about modularity just described are closely connected to the ideas about the graphical representation of interventions that were introduced in chapters 2 and 3. Recall that we represented an intervention on a variable  $X$  by drawing an arrow directed into  $X$  from the intervention variable and removing all other arrows directed into  $X$ . All other arrows in the graphical representation, including all arrows directed out of  $X$ , are preserved. The idea that arrows directed out of the variable intervened on are preserved represents a qualitative version of level invariance: the thought is that if a directed arrow from  $X$  to  $Y$  represents a genuine causal relationship, then that relationship, and hence the arrow, should be preserved under (some) interventions that change the value of  $X$ . The idea that all other arrows in the graph (i.e., those that are neither into or out of  $X$ ) are preserved is a qualitative version of the modularity requirement. As before, this may be motivated by the assumption that each set of arrows directed into a variable  $X$  represents a distinct mechanism that ought to be changeable independently of other mechanisms. If a graph (and the corresponding system of equations) is not modular, then the graph will exhibit action at a distance in the sense that it will behave as though interventions on some variables change arrows directed into or out of other variables. This indicates that the graph does not accurately and completely represent the causal relationships among those variables.<sup>9</sup>

## 7.5 Modularity in Other Contexts

So far, I have been discussing the notions of modularity and mechanism in the context of systems of structural equations. However, the analysis described

above is applicable much more generally. In a wide variety of cases, understanding the behavior of a complex system is a matter of representing it as segregated into parts or components, where the representation is modular in the sense that those components are represented as changeable independently of each other. As an illustration, return to the example from chapter 1 in which a block of mass  $m$  slides down an inclined plane under the influence of a gravitational field with a component of gravitational force  $F_g$  directed downward along the plane and a frictional force  $F_k$  that is proportional to the velocity of the block and opposed to its direction of motion. Gravity and friction are distinct forces, deriving from distinct physical mechanisms or relationships. If the representation of these forces that we have chosen is modular, it should be possible in principle to change the generalization governing the gravitational force on the block without changing the generalization governing the frictional force and vice versa. In fact, this is what we do find. For example, we might alter the relationship between the frictional force and the velocity of the block and hence the generalization that describes this relationship but not the relationship describing the influence of the gravitational field on the block by greasing the surface of the plane. This will change the coefficient of kinetic friction, but will not alter the generalization governing the gravitational force component. On the other hand, we could in principle alter the latter relationship by moving the inclined plane to a weaker gravitational field, which would result in a different value for the acceleration due to gravity. This would also change the value of  $N$ , the normal force exerted by the block, but it should not alter the relationship (1.3.2)  $F_k = \mu_k N$  between the frictional force  $F_k$  and  $N$ .

When we adopt the usual physical analysis of this example in which the total force along the plane on the block is decomposed into distinct components due to gravity and friction (as in 1.4.4), we implicitly respect these ideas about modularity and independent changeability. We think of the correct decomposition of a total force into components as just the decomposition in which the components can be changed independently of each other or in which the force law for each component is invariant under changes in the other component. Just as in Haavelmo's example, it will be this representation that will be best suited to tracing the results of hypothetical changes and answering what-if-things-had-been-different questions, hence, the most perspicuous representation from the point of view of explanation. By contrast, suppose that we represent the net force on the block as the sum

$$(7.5.1) \quad F_{net} = F_1 + F_2$$

where  $F_1$  and  $F_2$  are vectors conforming to

$$(7.5.2) \quad F_1 = aF_k + bF_g$$

$$(7.5.3) \quad F_2 = cF_k - dF_g \quad \text{with} \quad a + c = b - d = 1.$$

Here,  $F_1$  and  $F_2$  are not changeable independently of each other, which is to say that they do not describe the operation of distinct mechanisms: they are like (7.4.3)–(7.4.5) on the assumption that (7.4.3)–(7.4.4) is modular. If we are given the relationship (7.5.1) and are told that the value of  $F_1$  has changed, we must assume that  $F_2$  has changed as well, and without additional information we have no basis for saying how  $F_2$  will have changed.

Representations of the behavior of a system that are modular have another, closely related virtue: they allow us to combine representations of different systems to build up representations of yet more complex systems. Suppose that we add an additional component to the above system that consists of a rope connecting a second weight to the sliding block via a pulley. This will result in an additional force on the block. If we can assume, in accordance with modularity, that this addition will not alter the previous two force components due to friction and the gravitational force on the block or the relationships governing these, we can simply combine the original analysis of the block sliding down the plane with a generalization governing the behavior of the additional force component to produce an account of this new, more complex mechanism.

Many representations of biological structures also exhibit a similarly modular character. Consider the operation of lac operon described in chapter 6. The wild type gene  $I^+$  that produces the repressor protein can be replaced with a mutant form of the gene  $I^-$  that does not produce the repressor. As expected, cells containing only  $I^-$  but normal structural genes synthesize full levels of the enzymes in both the presence and the absence of an inducer. This is expected because, in accordance with modularity, it is assumed that this way of changing the repressor gene from  $I^+$  to  $I^-$  does not affect the generalizations describing the operation of the structural genes, although it does, of course, cause the structural genes to be continually “on.” If cells contain both  $I^-$  and a mutant form  $Z^-$  of the structural gene that in its unmutated form ( $Z^+$ ) synthesizes B-galactosidase, they will not produce this enzyme. If an  $I^+Z^+$  chromosome fragment is then introduced, recipient cells synthesize B-galactosidase for a period and then stop. The obvious interpretation is that this occurs because by the end of this period, enough repressor product has been produced by the  $I^+$  genes to block further synthesis.

Modularity is thus desirable (and achievable) not just in systems of equations in causal modeling contexts but more generally. I emphasize, however, that to say that it is desirable, from the point of view of explanation, to find a modular representation of the operation of complex systems whose behavior we are trying to understand is not to say that we are guaranteed to find such a representation at any given level of analysis. For example, it seems to me entirely possible that for many complex psychological processes no decomposition into components at the relatively coarse-grained level represented by the boxological diagrams found in cognitive psychology textbooks will be modular; instead, the only modular decomposition may be found at a much more fine-grained (e.g., neurobiological) level.

## 7.6 Invariance and Probabilistic Causation

My discussion so far has largely focused on examples involving linear equations, but it is readily generalized to equations involving arbitrary functional forms. By doing this we can also connect the ideas about invariance described above to an explicitly probabilistic framework for thinking about causation that may be more familiar to philosophers. Following Pearl (2000a), let us define a causal theory  $T$  as a four-tuple  $T = \langle V, U, P(u), \{f_i\} \rangle$  where

- (i)  $V = \{X_1 \dots X_n\}$  is a set of measured variables
- (ii)  $U = \{U_1 \dots U_n\}$  is a set of exogenous error variables
- (iii)  $Pr(u)$  is a probability distribution over  $U_1 \dots U_n$ .
- (iv)  $\{f_i\}$  is a set of  $n$  functions, each having the form  $X_i = f_i(\text{Parents } X_i, U_i)$  for  $i = 1 \dots n$ .

The probability distribution  $P(u)$  together with the functions  $f_i$  determine a probability distribution over all the variables in  $V$ . The underlying causal relationships are thus assumed to be deterministic, with the stochastic element in the model being supplied by the error variables  $U$ . (We will relax this assumption below).

Within this framework, to say that some individual equation  $X_i = f_i(\text{Parents } X_i, U_i)$  is level-invariant is just to say that the function  $f_i$  (whatever it may be) is invariant or continues to hold on some range of interventions on the variables  $\text{Parents } X_i$  and  $U_i$ . As before, we may think of each individual equation as representing a distinct causal mechanism. If mechanisms are distinct, then it ought to be possible to disrupt or change the corresponding equation without changing other equations in  $T$ . As before, if  $T$  meets this condition, we may describe it as modular or equation-invariant.

Again following Pearl and our brief remarks in chapter 2, let us introduce an operator *set* ( $X = x$ ) to represent the fact that the value of the variable  $X$  has been set equal to  $x$ . We may then represent the effect of an intervention which sets  $X_i = x$  in the following way: we delete (or “wipe out”) from  $T$  the equation  $f_i$  in which  $X_i$  is the dependent variable and replace it with equations that specify  $X_i = x$ . On the assumption that the system is modular, it will be possible to do this in such a way that all other equations are left undisturbed. By then substituting  $X_i = x$  for every occurrence of  $X_i$  among the independent variables (parents) in each of the  $f_i$ , we may trace out the effects of the intervention  $X_i = x$ . If we are willing to assume that *set*  $X$  is itself a random variable with a well-defined probability distribution (as would be the case if, for example, the values of *set*  $X$  are determined by some randomizing device), then we may talk about the probability distribution of some other variable  $Y$  conditional on  $X$ 's being set to various values, that is,  $Pr(Y/\text{set } X)$ .

Given this understanding of the “set” operator and its connection with the “wipe out” procedure for equations described above, various rules or requirements governing its behavior follow. For example, we have the following

probabilistic analogue to **MODULARITY** which we may call **PROBABILISTIC MODULARITY (PM)**:

**(PM)**  $Pr(X/Parents(X)) = Pr(X/Parents X \cdot set Z)$  where  $Z$  is any set of variables distinct from  $X$ .

**PM** expresses the idea that once one conditions on the full set of causes of  $X$ , setting any other variable should make no additional difference to the probability of  $X$ . Each conditional probability  $Pr(X_i/Parents X)$  is determined by the corresponding equation  $X_i = f_i(\text{parents } X_i, U_i)$  and the probability distribution over  $U_i$ . Hence, we may think of each conditional probability (like each equation) as corresponding to a distinct mechanism. **PM** is thus another way of expressing the idea that we may disrupt any other mechanism in the system (by setting the dependent variable for that mechanism equal to some exogenously determined value) without disrupting the conditional probability  $Pr(X/Parents X)$ . Note also that it will not in general be true that  $Pr(X/Parents X) = Pr(X/Parents X \cdot Z)$ ; that is, if one replaces *set Z* in **PM** with  $Z$ , the resulting statement will no longer be true. For example, conditioning on an effect  $Z$  of  $X$  may provide information about the value of  $X$  even when the values of the parents of  $X$  are taken into account. By contrast, if the value of the same variable  $Z$  is set by an intervention, then it will provide no information about the value of  $X$  because the effect of the intervention is to break the previously existing causal relationship between  $X$  and  $Z$ . This provides a further illustration of the difference between conditioning and intervening.

We also have an obvious probabilistic analogue of level invariance:

**PROBABILISTIC LEVEL INVARIANCE (PLI)**  $Pr(X/Parents X) = Pr(X/set Parents X)$ .

This represents the idea that if *Parents X* are the causes of  $X$ , then the conditional probability  $Pr(X/Parents X)$  should be invariant under interventions that change the value of any of the variables in *Parents(X)*. As before, this is likely to be true only for some range of interventions on *Parents(X)* and not for all such interventions.

If we are willing to assume in addition that the causal theory with which we are dealing satisfies the Causal Markov condition **CM** (which, it will be recalled, says that, conditional on its parents, every variable in  $V$  is independent of every other variable except its effects), then a number of additional rules governing the set operator will hold, again for some range of interventions. Two such rules, discussed in Hausman and Woodward (1999) and labeled by us **PM2** and **PM3** are:

**(PM2)** When  $X$  and  $Y$  are distinct,  $Pr(X/set Parents Y \cdot Y) = Pr(X/set Parents Y \cdot set Y)$ .

**(PM3)** If  $X$  does not cause  $Y$ , then  $Pr(X/Parents X \cdot set Y) = Pr(X/Parents X \cdot Y)$ .

**PM2** says that it should make no difference to the value of  $X$  whether we set  $Y$  or observe  $Y$ , once we set *Parents(Y)*. Moreover, this should be the case even if

$Y$  is a descendant or parent of  $X$ . **PM3** says that when  $X$  does not cause  $Y$ , then, conditional on *Parents*( $X$ ), *set*  $Y$  and  $Y$  should have the same probabilistic impact on  $X$ . As explained in Hausman and Woodward, we may think of the Causal Markov condition **CM** as the conjunction of **PM** and **PM3**. All of these rules—**PM**, **PLI**, **CM**, **PM2**, and **PM3**—capture different invariance conditions that it may be reasonable to impose in probabilistic contexts. They illustrate how different claims about invariance can be stated with some precision within such contexts.<sup>10</sup>

So far, I have been assuming a framework in which causal relationships are deterministic or, more accurately, pseudo-indeterministic. Do invariance conditions like those described above also hold in contexts that are irreducibly stochastic? That depends on how we think probabilistic causes behave. A very natural assumption<sup>11</sup> frequently made in philosophical discussions is that, putting aside the values taken by the effects of a variable  $X$ , the values taken by the full set of direct causes of  $X$  fully determines the probability that  $X$  will assume the various values in its range. (This is an implication of the Causal Markov condition (**CM**)). Given **CM**, one may think of each conditional probability  $Pr(X/\textit{parents } X)$  as corresponding to a distinct stochastic mechanism or causal relationship, with different mechanisms being changeable or disruptable independently of each other. In this sort of framework, it is natural to expect invariance conditions like **PM** and **PLI**, **PM2**, and **PM3** to hold. Suppose, for example, that  $C$  is the only parent (i.e., the only direct cause) of  $X$  and  $Y$ , with  $C$ ,  $X$ , and  $Y$  all being dichotomous variables. Then, corresponding to the causal relationship between  $C$  and  $X$  will be a characteristic conditional probability  $Pr(X/C)$ , and corresponding to the relationship between  $C$  and  $Y$  will be another characteristic conditional probability  $Pr(Y/C)$ . A change in  $P(C)$  (e.g., an increase in the number of cases in which  $C = 1$ ) will of course change  $Pr(X)$  and  $Pr(Y)$ , but (at least within a certain range of changes in  $Pr(C)$ ) should not change the conditional probabilities  $Pr(X/C)$  and  $Pr(Y/C)$ . These should be stable or invariant under such changes. This is the sort of invariance captured by **PLI**. Similarly, it should be possible to intervene on  $Y$  and change the relationship between  $C$  and  $Y$  without altering  $Pr(X/C)$ , and this idea is captured by **PM**. If instead we had claimed that  $Pr(C/XY)$  was invariant under interventions on  $X$  and  $Y$ , this would be a way of encoding or representing a different set of claims about causal structure: that  $X$  and  $Y$  are causes of  $C$  and that  $Pr(C/XY)$  represents a distinct causal mechanism.

It is important to distinguish the claim that a probability distribution can be factored in a certain way from an invariance condition like **PM**. **PM** says instead that certain terms (and the variables or mechanisms corresponding to them) can be changed without disrupting other terms (variables and mechanisms). Any probability distribution can be factored in many different ways, but at most one of these factorizations will consist of terms, any one of which may be changed independently of the other terms in accordance with **PM**. As an illustration, consider that the joint distribution  $Pr(A, B)$  can be written as either  $Pr(A) \cdot Pr(B/A)$  or as  $Pr(B) \cdot Pr(A/B)$ . However, this fact does not show which if either of these factorizations consists of terms that may be changed independently

of the others. If, for example,  $A$  is the sole cause of  $B$ , then from **PM** and **PLI** one should expect that  $Pr(B/A)$  will be stable under (some range of) interventions that change  $Pr(A)$ . Similarly, if  $B$  is the sole cause of  $A$ , then  $Pr(A/B)$  should be stable under interventions on  $Pr(B)$ . Moreover, it is easy to show that if  $Pr(B/A)$  is nonzero for both values of  $A$  and invariant to interventions that set  $A$ , then the inverse conditional probability  $Pr(A/B)$  cannot be invariant. Changing notation so that  $\{A, -A\}$  now represent the values of  $A$ , we see from Bayes's theorem that  $Pr(A/B) = [Pr(A) \cdot Pr(B/A)] / [Pr(A) \cdot Pr(B/A) + Pr(-A) \cdot Pr(B/-A)]$ . If  $Pr(B/A)$  and  $Pr(B/-A)$  are invariant under interventions that change  $Pr(A)$ , then  $Pr(A/B)$  must change for different values of  $Pr(A)$ . Alternatively, if it is not true that either  $A$  causes  $B$  or that  $B$  causes  $A$ , then an intervention that changes the value of either leaves the distribution of the other unchanged and hence disrupts the conditional probabilities  $Pr(A/B)$  and  $Pr(B/A)$ . (For a similar argument, see Sober 1994, pp. 234–37; Hoover 2001, chap. 8.)

## 7.7 Causal Information as Inputs to Structural Models

My discussion so far has focused almost entirely on issues of causal interpretation. I have said little about epistemological issues concerning how causal claims of the sort embodied in structural models may be established or about the kinds of casual information that typically figures as inputs to structural models.

There is very general agreement that causal information of some kind is required as input or as background assumptions if causal modeling techniques are to yield reliable causal conclusions: the causal claims embodied in structural models do not emerge just from statistical data, but from such data in conjunction with additional assumptions which are “extra statistical” in the sense that their justification does not come just from the statistical data at hand (i.e., from information about population variances and covariances among the measured variables) but rather has some other rationale. The need for such additional assumptions should be apparent from the observations about empirical equivalence and underdetermination made above: often, a large number of different causal models will be consistent with the observed statistical data, and hence to discriminate among these additional constraints or assumptions are required.

In principle, these additional assumptions might take either of two different forms. On the one hand, they might be domain- (or subject-matter-) specific in the sense that they embody very specific empirical claims about, for example, the ability of one causal factor to influence another. Thus, if one knew on independent grounds that (7.7.1) a certain fertilizer will not influence the growth of some variety of plant, this would be a basis for excluding the former variable from the r.h.s. of a regression equation designed to establish the causes of plant height. This exclusion would count as one example of an extra-statistical assumption, that (as we shall see below) with the right combination of other assumptions and statistical evidence might be used to establish the

conclusion that certain other causes do influence plant height. On the other hand, the extra-statistical assumptions on which one relies might instead be relatively general and domain-independent. The Causal Markov and Faithfulness conditions, discussed in chapters 2 and 3, are examples of such principles.

Both kinds of assumptions (general or specific) are causal in character in the sense that causal notions apparently occur in them in an unreduced or primitive form. This is obvious in the case of domain-specific assumptions like (7.7.1), but it is equally true of more general assumptions like the Causal Markov and Faithfulness conditions. Not only do these employ explicitly causal locutions in their formulation, but it is also true, as I have argued elsewhere (Woodward 1997a; Hausman and Woodward 1999), that their reliable use in causal inference requires supplementation by domain-specific causal information. Both because of this fact and because I have discussed the strengths and limitations of the use of domain-general principles in causal inference elsewhere (Woodward 1997a, 1998), I focus in what follows on the use of domain-specific information.

One variety of domain-specific causal information that plays an important role in establishing many structural models is what I called *possible-cause* or *capacity* information in chapters 5 and 6: information that one variable “can cause” or has the capacity to cause another. One role for such information is to justify claims about which variables should be included or excluded from a structural model: one argues for the inclusion (or exclusion) of a variable  $X$  as an independent variable in a structural equation with dependent variable  $Y$  by showing that  $X$  has (or does not have) the capacity to causally influence  $Y$  in a range of background conditions that include those obtaining in the population on which the structural model is estimated. Such capacity information plays other roles as well; for example, it can be used to justify claims about the uncorrelatedness of the error term. In my view, the notion of capacity that is relevant to structural models is just the notion described in chapter 6: when  $X$  has the capacity to cause  $Y$ , the relationship between  $X$  and  $Y$  will be (at least) weakly invariant in the sense that, for at least some individuals in various populations and in a range of background conditions, interventions that alter whether they possess some value of  $X$  will alter the value of  $Y$  they possess, even though this may not be true in all populations and background conditions and even though we may be unable to formulate any exact quantitative law linking  $X$  and  $Y$ .

To motivate the need for such capacity information in inferences involving structural models, consider the expression for the OLS estimator (7.1.8) for the coefficients in a regression equation. As (7.1.8) makes clear, the estimated value for each coefficient  $B_i$  will be a function not just of the covariance between  $X_i$  itself and the dependent variable  $Y$ , but also a function of (i) the covariance between  $X_i$  and all of the other independent variables in the equation and (ii) the covariance between these variables and  $Y$ . It follows that one can always alter the coefficient of any variable in a regression equation by the inclusion or deletion of other variables as long as these variables exhibit



a nonzero covariance with the other variables included in the equation and with  $Y$ . For example, in the two-variable multiple regression equation

$$(7.7.2) \quad Y = B_1 X_1 + B_2 X_2 + U$$

the OLS estimator for the coefficient  $B_1$  is

$$(7.7.3) \quad B_1^* = [S_{x_2}^2 \cdot S_{yx_1} - S_{yx_2} S_{x_1 x_2}] / [S_{x_1}^2 S_{x_2}^2 - (S_{x_1 x_2})^2]$$

where  $S_{x_2}^2$  is the sample variance of  $X_2$ ,  $S_{yx_1}$  the sample covariance between  $Y$  and  $X_1$ , and so on. Hence, the estimated value of the coefficient  $B_1$  depends on the correlation between  $Y$  and  $X_1$ , the correlation between  $X_1$  and  $X_2$ , and the correlation between  $X_2$  and  $Y$ . As long as  $X_2$  exhibits a nonzero correlation with  $X_1$  and  $Y$ , we can change the estimated value of  $B_1$  by dropping  $X_2$  from the equation or by substituting for  $X_2$  a new variable  $X_3$  which has a different correlation with  $X_1$  and  $Y$ . In many realistic cases, it will be easy to find a large number of different variables in one's data set (or in the population of interest) that are correlated with  $Y$  and with the other independent variables, although they may not be causally relevant to  $Y$ . Without additional constraints on which variables it is appropriate to include in the equation, there will be no reason to prefer one set of coefficients to any other. When it is available, information about causal capacities can provide such constraints.

An illustration of this role for capacity information is provided in Tufte (1974), which analyzes the causes of variations in automobile fatality rates across states. Although Tufte's analysis relies mainly on simple scatter plots and tabular comparisons, we can translate it into a multiple regression context. Thinking about the problem in this way, what Tufte in effect does is to regress a variable representing death rates for the United States against a number of other independent variables, including a variable representing presence or absence of automobile inspections, a variable representing population density, and variables representing whether or not a state was one of the original thirteen states and has seven or fewer letters in its name. He obtains a nonzero regression coefficient in each case: death rate is correlated with each of these variables and, as it happens, each variable is correlated with the others. Tufte takes it to be obvious that if the regression equation is to be interpreted causally, inspections and population density are appropriate variables to include as independent variables, whereas other variables described above, such as number of letters in a state's name, are not. This view is not based on the pattern of statistical association among these variables and the death rate but rather on extrastatistical considerations, which have to do with what Tufte calls "substantive judgment." He writes: "While we observe many different associations between the death rate and other characteristics of the state, it is our substantive judgment, and not merely the observed association that tells us density and inspections might have something to do with the death rate and that the number of letters in the name of the state has nothing to do with it" (p. 9).

My suggested reconstruction of these remarks is that the judgment that the presence of automobile inspections “might have something to do with” fatality rates can be seen as based on the claim that it is plausible that inspections have the capacity to influence fatality rates in the sense that there is a range of circumstances in which there will be a pattern of association between whether inspections occur and the fatality rate that is weakly invariant. Although we might well wonder what the quantitative impact of inspections generally or of a particular program of inspections is on fatalities—whether the impact is large or small or cost-effective—there is general agreement that inspections are “the sort of thing” that “can affect” fatalities in a range of circumstances; indeed, inspections are introduced for just this reason. The reasons for thinking that population density has the capacity to affect fatality rates are somewhat less transparent, but Tufté is able to show that this is just what we would expect, given other widely shared causal assumptions. Thinly populated states have higher fatality rates compared to thickly populated states because drivers go for longer distances at higher speeds, accidents are more likely to be severe when they occur at a higher speeds, and victims are less likely to be discovered and treated immediately (pp. 20–21).

In contrast to the situation with inspections and population density, our background causal knowledge insists that the number of letters in a state’s name does not have the capacity to (is not the sort of thing that could) causally influence the automobile fatality rate. No one seriously supposes that there are realistic circumstances in which shortening state names would serve as an effective highway safety measure, and for this reason we do not regard it as an appropriate variable to enter into a regression equation in which highway fatalities is a dependent variable.

Although extra-statistical causal knowledge about causal capacities is used to reach causal conclusions in the above examples, if the argument of chapter 6 is correct it is misguided to think of this as knowledge of laws of nature. Claims like (7.7.4) “Automobile inspections can influence the traffic fatality rate” are far too vague, imprecise, and exception-ridden to be assimilated to laws. Although (7.7.4) is not contentless—it tells us that there is some range of circumstances in which manipulating whether inspections occur is a way of altering the fatality rate—it does not tell us with any precision exactly when inspections will influence fatalities and by how much. Moreover, the features we ordinarily associate with laws of nature, such as exceptionlessness (at least within a certain domain) and precision, are not required if capacity claims are to play the role of input to structural models. For the claim that  $X$  has the capacity to cause  $Y$  to play the role described above, we don’t need to know an exceptionless generalization linking  $X$  to  $Y$  or exactly how  $X$  influences  $Y$  in all possible background circumstances. Instead, it is enough that we know that there is some range of circumstances in which  $X$  will or may produce this effect and that individuals in the population for which we are estimating the structural model fall into these circumstances. Structural models thus provide an additional illustration that what I called the epistemological thesis in

chapter 4—the thesis that reliance on laws is always required to establish causal claims—is false.

## 7.8 Structural Models as Expressing Population-Specific Causal Truths

I have been focusing on the causal information that figures as inputs to structural models. What about the structural models themselves? What sort of causal knowledge do they represent? When we estimate a structural model, our interest is typically not (or not just) in establishing a capacity claim, but rather in establishing a quantitative generalization about the impact of changes in one or more independent variables on a dependent variable in some particular population. In contrast to capacity claims, which are typically qualitative, the coefficients in structural models are quantitative: they purport to tell us exactly how much difference a change in this or that variable will make to some effect of interest in a particular population. The difference between a capacity claim and the causal claim represented in a structural model is thus roughly the difference between the claim that some cause *can* produce some effect in populations like the population of interest and the claim that it actually is at work in that population producing an effect of a certain magnitude. The claim that smoking can cause lung cancer is a capacity claim and a claim to which we might appeal to justify the inclusion of a variable measuring smoking behavior in a structural model designed to determine the factors that influence the incidence of lung cancer in the population of the contemporary United States. By contrast, the claim that such and such a change in per capita cigarette consumption will result in such and such a change in the incidence of lung cancer in the U.S. population is the sort of claim that we might hope to establish by means of structural equation techniques.

The researcher's intent in estimating a structural equation is to arrive at a quantitatively precise relationship with particular numerical values for the estimated coefficients; however, it is a matter of empirical fact that in economics and the other social sciences these quantitative relationships often hold only for the population or particular background circumstances for which they are estimated and not for other populations or background circumstances. That is, whereas qualitative capacity claims are weakly invariant across different populations or significant changes in background conditions, the quantitative coefficients in structural models often are not stable in the same way. For this reason, among others, we should not think of the structural equations themselves as laws.

As an illustration, consider the experiments carried out in different U.S. states in the 1970s to test the effects of a negative income tax or income maintenance program on labor market participation (Stafford 1985). In virtually all of these populations, there is a broadly similar qualitative pattern: the negative income tax has a relatively small effect on the labor market response of primary wage earners and a considerably more substantial effect on

secondary wage earnings. This is the sort of evidence that supports claims about the capacity of the negative income tax to produce these effects. Nonetheless, the coefficients representing the quantitative values of these effects vary in nontrivial ways across the different populations studied. For example, estimates of uncompensated wage elasticity for the entire population of U.S. adult males range from  $-0.19$  to  $-0.07$  and estimates of income elasticity range from  $-0.29$  to  $0.17$ . Estimates of uncompensated wage elasticity from two negative income experiments (Gary and Seattle-Denver) yield values of 0 (Gary),  $-0.06$  (short-run, Seattle-Denver), and  $0.02$  (long-run, Seattle-Denver). Estimates of income elasticity from these two experiments yield values of  $-0.05$  and  $-0.08$ , respectively. Labor market response also varies significantly for members of different ethnic groups and varies as well with marital status and gender (for discussion, see Stafford 1985, especially pp. 104–12).

On reflection, this variability is not surprising. Plainly, the effect of the negative income tax on labor market participation itself depends on a number of background characteristics of the population under study: these have to do with general attitudes toward work, the availability of other social welfare services, the availability of various employment opportunities, the existence of certain differentiated sex roles, and so on. Such characteristics vary somewhat across the different populations studied, and, for a variety of reasons, it is not always possible to control for them by entering them explicitly into one's model as additional variables. We thus see, as argued above, a contrast between qualitative capacity claims about the effects of the negative income tax on primary and secondary wage earners and on divorce rates, which do hold across a range of different populations, and the more precise quantitative claims embodied in structural models which are far more population-specific. Nonetheless, despite their population specificity, there is no bar, on the account that I have been sketching, to regarding such quantitative relationships as expressing genuine causal truths: they tell us, for example, about the quantitative effect on wages of (some range of) interventions that set the negative income tax at various levels in particular populations, such as the people living in Gary or Seattle.

Similar conclusions follow for other examples we have considered. I noted that Veblen (1975; discussed in chapter 5) estimated that a move from relatively neutral to highly favorable coverage by the *Manchester Union Leader* in a New Hampshire primary election changes a candidate's vote by 19 percent and that in a general election the difference is 14 percent (cf. (5.6.1)). However, both Veblen and Achen (1982, commenting on Veblen) make it explicit that these are results about the causal influence of a particular newspaper in a particular population during a particular time period. It would be absurd to suppose that these quantitative estimates can play the role of constants in a universal law concerning the influence of newspapers on election results. Achen recognizes this when he asserts elsewhere that the researcher's intent, in using regression analysis, is to describe, for example, "the effect of the Catholic vote on the Nazi rise to power or the impact of a preschool cultural enrichment program like Head Start on poor children's success in school. Whatever the truth in such cases, one would not characterize it as a law.

Neither Catholics nor impoverished youngsters would behave the same way in other times and places" (p. 12).

On the view that I have been defending, the population specificity of the relationship that Veblen estimates is no bar, in itself, to regarding it as causal: the relationship will qualify as causal as long as it is true that, in the New Hampshire electorate during the period 1960–1972, there is some range of interventions that alter the editorial slant of the *Union Leader* and will change the vote by the amounts described above.

The variability and instability of the coefficients in typical structural models is acknowledged by many other researchers. Thomas Mayer (1980) describes a number of empirical studies of consumption functions, investment functions, money supply, and demand functions, all of which seem to suggest considerable coefficient instability. Although this instability derives in some cases from correctable methodological lapses, Mayer acknowledges that in many cases, it results from what he describes as the "more general problem that behavioral parameters are not as stable as those in the natural sciences" (p. 168). Failure of stability is also acknowledged in many methodological treatises. For example, Newbold and Bos (1985) write:

There is a crucial distinction between much physical scientific and social scientific theory. In the physical sciences, theory often suggests the presence of physical constants, yet only rarely is this truly the case in the social sciences. For example, we know that if a heavy object is dropped to the ground today in Champaign, Illinois, it will accelerate at a rate of approximately 32 feet per second. Moreover, the rate of acceleration would be very nearly the same in Melbourne, Australia, today or in Champaign, Illinois in twenty years' time. It is very difficult to think of any social scientific theory that strongly asserts the existence of such a physical constant. In economics, theories of the consumption function and of consumer demand are well developed. It is not, however, asserted as a result of such theories, that the marginal propensity to consume is the same in the United States as in Australia, or that the elasticity of demand for coconuts is the same in Champaign as in Melbourne. (p. 11)

In a similar vein, Johnston (1992) writes in an overview of recent work in econometrics:

One impression which surfaces repeatedly in any perusal of applied work is the fragility of estimated relationships. By and large these relationships appear to be time specific, data specific and country or region specific. Should one expect, for example, a stable demand function for apples? If so, would one expect that function to encompass the behavior of Eve in the Garden of Eden and that of her present-day sisters, such as the Newport Beach Matrons strolling the aisles of the upscale super markets in Southern California? The question need only be posed to provide the answer. (p. 53)

The contrast between physical laws and population-specific structural equations described in these passages dovetails with my insistence in earlier chapters that it is unilluminating to assimilate all general causal relationships to laws of

nature. The appeal of this assimilation derives from the tacit assumption that either a generalization expresses a law or else it describes a merely accidental noncausal correlation. The ideas about invariance described above provide us with a way of understanding how structural equations can express local causal truths without qualifying as laws.

## The Causal Mechanical and Unificationist Models of Explanation

This chapter explores the relationship between the account of causation and explanation developed in previous chapters and two of the most influential alternative accounts: the *causal mechanical* (CM) model developed by Wesley Salmon and the *unificationist* model developed by Philip Kitcher. I follow the usual philosophical convention of focusing on (what I take to be) the comparative advantages of the manipulationist account. Despite its critical character, I hope that it will be clear from my discussion that both the CM and unificationist models are valuable philosophical achievements from which there is much to be learned.

### 8.1 The CM Model and Causal Relevance

The CM model employs several central ideas. A *causal process* is a physical process, such as the movement of a baseball through space, that is characterized by the ability to transmit a *mark* in a continuous way. (“Continuous” generally, although perhaps not always, means “spatiotemporally continuous.”) Intuitively, a mark is some local modification to the structure of a process: a scuff on the surface of a baseball or a dent in an automobile fender. A process is capable of transmitting a mark if, once the mark is introduced at one spatiotemporal location, it will persist to other spatiotemporal locations even in the absence of any further interaction. In this sense, the baseball will transmit the scuff mark from one location to another. Similarly, a moving automobile is a causal process because a mark in the form of a dent in a fender will be transmitted by this process from one spatiotemporal location to another. Causal processes contrast with *pseudo-processes*, which lack the ability to transmit marks. An example is the shadow of a moving physical object. The intuitive idea is that, if we try to mark the shadow by modifying its shape at one point (e.g., by altering a light source or introducing a second occluding object), this modification will not persist unless we continually intervene to maintain it as the shadow occupies successive spatiotemporal positions. In other words, the modification will not be transmitted by the structure of the shadow itself, as it would in the case of a genuine causal process.

We should note for future reference that, as characterized by Salmon, the ability to transmit a mark is clearly a counterfactual notion, in several senses. To begin with, a process may be a causal process even if it does not in fact transmit any mark, as long as it is true that *if it were marked*, it would transmit the mark. Moreover, the notion of marking itself involves a counterfactual contrast: a contrast between how a process behaves when marked and how it would behave if left unmarked.

The other major element in Salmon's model is the notion of a *causal interaction*. A causal interaction involves a spatiotemporal intersection between two causal processes which modifies the structure of both: each process comes to have features it would not have had in the absence of the interaction. A collision between two cars that dents both is a paradigmatic causal interaction. Here too, the counterfactual element in this characterization is obvious.

According to the *CM* model, an explanation of some event *E* will trace the causal processes and interactions leading up to *E* (Salmon calls this the *etiological* aspect of the explanation), or at least some portion of these, as well as describing the processes and interactions that make up the event itself (the *constitutive* aspect of explanation). In this way, the explanation shows how *E* "fit[s] into a causal nexus" (1984, p. 9).

The suggestion that explanation involves "fitting" an explanandum into a causal nexus does not give us any very precise characterization of what the relationship between *E* and other causal processes and interactions must be if information about the latter is to explain *E* (a point to which I will return). Nonetheless, I believe that the following example illustrates what Salmon has in mind.

Suppose that a cue ball, set in motion by the impact of a cue stick, strikes a stationary 8 ball with the result that the 8 ball is put in motion and the cue ball changes direction. The impact of the stick also transmits some blue chalk to the cue ball, which is then transferred to the 8 ball on impact. The cue stick, the cue ball, and the 8 ball are causal processes, and the collision of the cue stick with the cue ball and the collision of the cue and 8 balls are causal interactions. Salmon's idea is that citing such facts about processes and interactions explains the motion of the balls after the collision; by contrast, if one of these balls casts a shadow that moves across the other, this will be causally and explanatorily irrelevant to its subsequent motion because the shadow is a pseudo-process.

However, as Christopher Hitchcock (1995) shows in an illuminating paper, the information about causal processes and interactions just described leaves out something important. Consider the usual elementary textbook "scientific explanation" (call this 8.1.1) of the motion of the balls following their collision. This proceeds by deriving that motion from information about their masses and velocity before the collision, the assumption that the collision is perfectly elastic, and the law of the conservation of linear momentum. On the manipulationist conception of explanation, we should think of this derivation as explanatory because it systematically answers a set of *w-questions* about how the subsequent motion of the balls would have changed had the masses and



initial velocities of the balls been different. It is the exhibition of this pattern of counterfactual dependence that allows us to see that it is the mass and velocity of the balls, rather than, say, their color or the presence of the blue chalk mark, that are explanatorily relevant to their subsequent motion. However, it is hard to see what in the *CM* model allows us to pick out the linear momentum of the balls, as opposed to these other features, as explanatorily relevant. Part of the difficulty is that to express such relatively fine-grained judgments of explanatory relevance (that it is linear momentum rather than chalk marks that matters), we need to talk about relationships between properties or magnitudes, and it is not clear how to express such judgments in terms of facts about causal processes and interactions. Both the linear momentum and the chalk mark communicated to the cue ball by the cue stick are marks transmitted by the spatiotemporally continuous causal process consisting of the motion of the cue ball, and are then transmitted via an interaction to the 8 ball. In other words, there appears to be nothing in Salmon's notion of mark transmission or the notion of a causal process that allows one to distinguish between the explanatorily relevant momentum and the explanatorily irrelevant blue chalk mark: both count as marks transmitted by the ball. To capture this distinction, it looks as though we must appeal to counterfactuals—and counterfactuals that are different from those used to characterize mark transmission. This is what the manipulationist account does. The blue mark is explanatorily irrelevant because the appropriate pattern of counterfactual dependence is absent: the subsequent motion of the ball would have been no different whether or not the mark was present.

Ironically, as Hitchcock goes on to note, a similar observation may be made about a number of the counterexamples (discussed in chapters 4 and 5 above) originally devised by Salmon to illustrate the failure of the *DN* model to capture the notion of explanatory relevance. Spatiotemporally continuous causal processes that transmit marks as well as causal interactions are at work when Mr. Jones ingests birth control pills: the pills dissolve, components enter his bloodstream, are metabolized or processed in some way, and so on. Similarly, causal processes (albeit different processes) and spatiotemporally continuous paths are at work when Ms. Jones takes birth control pills. Intuitively, it looks as though the relevance or irrelevance of the birth control pills does not just have to do with whether the actual processes that lead up to Mr. Jones's non-pregnancy are capable of mark transmission but rather, as argued in chapter 5, (roughly) with the contrast between what happens in the actual situation in which Jones takes the pills and an alternative situation in which Mr./Ms. Jones does not take the pills. It is because the outcome (nonpregnancy) would be the same in both cases if Jones is male that the pills are explanatorily irrelevant. And it is because the outcome would not or might not be the same in both cases if Jones is a women that the pills are relevant in her case.

Similarly, consider the following variant on an example originally due to Kyburg (1965). A sample of salt is "hexed" by the touch of a witch's wand and then dissolved in water. Although there are spatiotemporally continuous processes that are capable of mark transmission leading from the hexing (or

the touch of the wand) to the dissolving of the salt, the hexing is explanatorily irrelevant to the dissolving. The challenge is to capture this judgment within the *CM* framework. Again, it looks as though counterfactuals are required to characterize the notions of causal and explanatory relevance and that one cannot characterize these notions just in terms of mark transmission and/or purely geometrical notions like spatiotemporal continuity.

A more general way of putting the problem is that those features of a process *P* in virtue of which it qualifies as causal (ability to transmit mark *M*) may not be the features of *P* that are causally or explanatorily relevant to the outcome *E* that we want to explain (*M* may be irrelevant to *E*, which is instead caused by some other property *R* of *P*). Even if the mark transmission criterion correctly distinguishes between *causal processes* and *pseudo-processes*, it does not, as it stands, provide the resources for distinguishing those *features* or *properties* of a causal process that are causally or explanatorily relevant to an outcome and those features that are irrelevant.

There is an additional point that is worth making in connection with examples like the explanation (8.1.1) involving the colliding billiard balls. Chapter 5 noted that deductive or derivational structure plays an important role in some explanations. (8.1.1) illustrates one of the advantages of constructing an explicit derivation: it can be a particularly effective way of conveying fine-grained and detailed information about dependency relations. This advantage is particularly salient when we compare the derivation of the motion of the balls from the basic principle of the conservation of linear momentum with Salmon's treatment. Part of the problem with neglecting deductive structures in favor of looser notions like "fitting" an event into a causal structure or "nexus" is that the latter notion isn't fine-grained or discriminating enough: it doesn't identify in detail the conditions or properties on which the explanandum depends or the precise pattern of dependence. This is of central importance in successful explanation.

## 8.2 Explanations That Do Not Involve Explicit Tracing of Causal Processes

As noted above, the *CM* model assigns a central role in explanation to the tracing of spatiotemporally continuous causal processes. The model thus runs into difficulty with cases in which we seem to have explanations that do not trace such processes. Many examples have been described in previous chapters. There are examples, such as the explanations provided by Newtonian gravitational theory, that involve "action at a distance" in a physically interesting sense. There are explanations, such as those involving causation by omission or by double prevention, that do not involve a physically interesting form of action at a distance, but are nonetheless cases of causal connection without intervening spatiotemporally continuous processes or transfer of energy and momentum from cause to effect. Finally, there are explanations, some of which are discussed in more detail below, that do not explicitly cite spatiotemporally

continuous causal processes involving transfer of energy and momentum, even though we may think that such processes are at work at a more “underlying” level. Most explanations in disciplines like psychology and economics have this character. In all such cases, the manipulationist theory seems to do a better job than the *CM* model of reconstructing how such explanations work.

### 8.3 The *CM* Model and Complex Systems

A third, not unrelated set of worries has to do with how we are to apply the *CM* model to more complex systems that involve a large number of interactions among what, from a fine-grained level of analysis, are distinct causal processes. Suppose that a mole of gas is confined to a container of volume  $V_1$  at pressure  $P_1$  and temperature  $T_1$ . The gas is then allowed to expand isothermally into a larger container of volume  $V_2$ . One standard way of explaining the behavior of the gas—its rate of diffusion and its subsequent equilibrium pressure  $P_2$ —appeals to the generalizations of phenomenological thermodynamics (e.g., the ideal gas law, Graham’s law of diffusion, etc.). Salmon appears to regard putative explanations based on at least the first of these generalizations as not explanatory because they do not trace continuous causal processes; he thinks of the individual molecules as causal processes but not the gas as a whole.<sup>1</sup> However, as argued in chapter 5, it is plainly impossible to trace the causal processes and interactions represented by each of the  $6 \times 10^{23}$  molecules making up the gas and the successive interactions (collisions) it undergoes with every other molecule. The usual statistical mechanical treatment, which Salmon presumably would regard as explanatory, does not attempt to do this. Instead, it makes certain general assumptions about the distribution of molecular velocities and the forces involved in molecular collisions and then uses these, in conjunction with the laws of mechanics, to derive and solve a differential equation (the Boltzmann transport equation) describing the overall behavior of the gas. This treatment abstracts radically from the details of the causal processes involving particular individual molecules and instead focuses on identifying higher-level variables that aggregate over many individual causal processes and that figure in general patterns that govern the behavior of the gas.

This example raises a number of questions. Just what does the *CM* model require in the case of complex systems in which we cannot trace individual causal processes, at least at a fine-grained level? How exactly does the causal mechanical model avoid the (disastrous) conclusion that any successful explanation of the behavior of the gas must trace the trajectories of individual molecules? Does the statistical mechanical explanation described above successfully trace causal processes and interactions or specify a causal mechanism in the sense demanded by the *CM* model, and if so, what exactly does tracing causal processes and interactions involve or amount to in connection with such a system? As matters now stand, the *CM* model is incomplete. If it is to be a general theory of explanation, we need answers to such questions.

There is another aspect of this example that is worthy of comment. I argued in chapter 5 that even if, per impossible, an account that traced individual molecular trajectories were to be produced, there are important respects in which it would not provide the explanation of the macroscopic behavior of the gas that we are looking for—and not just because such an account would be far too complex to be followed by a human mind. This is because tracing the actual sequence of molecular trajectories and collisions that lead up to the final pressure  $P_2$  of the gas does not tell us about the full range of (counterfactual and not just actual obtaining) conditions under which  $P_2$  would have occurred and the (counterfactual) conditions under which it would have been different; that is, which other trajectories would or would not have produced pressure  $P_2$ . By contrast, the account of explanation that I have proposed captures the idea that this counterfactual information is relevant in a natural way.

A similar point seems to hold for explanations of the behavior of other sorts of complex systems, such as those studied in biology and economics. Chapter 5 made the point that descriptions of the behavior of such systems at the level of basic physics will contain a great deal of irrelevant information and may well omit relevant information. A related point holds for the levels of description of the causal processes and interactions at work in these systems that seemed to be demanded by Salmon's account, with its emphasis on energy and momentum transfer and spatiotemporal continuity. Consider again the explanation of the behavior of some market or economic entity such as (8.1.2) the increase in the price of oranges following a freeze or the output-restricting behavior of a monopolistic firm described in (5.1.1). Underlying the behavior of these systems are individual spatiotemporally continuous causal processes and interactions in Salmon's sense: there are myriad individual transactions in which money in some form is exchanged for physical goods, all of which involve transfers of matter or energy; there is exchange of information about intentions or commitments to buy or sell at various prices, all of which must take place in some physical medium and involve transfers of energy; and so on. However, it also seems plain that producing a full description of these processes (supposing, for the sake of argument, that it was possible to do this) will produce little or no insight into why these systems behave as they do. Again, this is not just because any such "explanation" will overwhelm our information-processing abilities. It is also the case that a great deal of the information contained in such a description will be irrelevant to the behavior we are trying to explain, for the same reason that a description of the individual molecular trajectories will contain information that is irrelevant to the behavior of the gas. For example, although the detailed description of the individual causal processes involved in the operation of the market for oranges presumably will describe whether individual consumers purchase oranges by cash, check, or credit card, whether information about the freeze is communicated by telephone or e-mail, and so on, all of this is to a first approximation irrelevant to the equilibrium price: given the supply and demand curves, the equilibrium price will be the same as long as there is a market in which consumers are able to purchase oranges by some means, information about the freeze and about

prices is available to buyers and sellers in some form, and so on. Whether cash or credit cards are used is irrelevant in just the sense captured by the manipulationist account; changes in what form of money is used make no difference to the outcome we are trying to explain. Moreover, those factors that *are* explanatorily relevant to the equilibrium price, such as the shape of the demand and supply curves, are not in any obvious sense themselves connected by spatiotemporally continuous processes to the price (it is unclear what this claim even means), although, as emphasized above, the unknown processes underlying the attainment of equilibrium are presumably spatiotemporally continuous.

Again, the issue is how an account like Salmon's can capture this feature of successful explanation of the behavior of complex systems—how the account guides us to find the “right” level of description of the phenomena we are trying to explain. In fact, as the above discussion illustrates, the requirements that Salmon imposes on causal processes, and in particular the requirement of spatiotemporal continuity, often seem to lead us away from the right level of description. Despite the fact that they are successful explanations, there is no obvious sense in which (5.1.1) the explanation of the behavior of the monopoly and (8.1.2) the shift in the price of oranges due to the freeze trace spatiotemporally continuous processes. Moreover, the level at which the spatiotemporal continuity constraint is most obviously respected (the level at which, e.g., we describe a particular consumer as exchanging cash for oranges or a grower as making an agreement via telephone with a retailer to sell at a certain price) seems to be the wrong level for achieving understanding. A theory like the one I have been defending that takes successful explanations to involve the exhibition of dependency relations and recognizes that there can be invariant relationships at different levels of description seems much better suited to capturing what is going on in such explanations.

#### 8.4 Causal Explanation without Counterfactuals?

In fact, however, in more recent work (e.g., Salmon 1994), Salmon drew exactly the opposite conclusion. In response to counterexamples advanced by Philip Kitcher (1989), among others, to the mark criterion and to his characterization of causal interactions, Salmon concluded that the problems with his original theory were due to its reliance on counterfactuals, for example, in the characterization of mark transmission. He accordingly attempted to fashion a theory that completely avoids any appeal to counterfactuals. I first comment briefly on this new theory and then on Salmon's reasons for abandoning the mark criterion.

Salmon's new theory, which is influenced by Dowe (see, e.g., Dowe 2000), characterizes the notion of a causal process in terms of the notion of a *conserved quantity*. A causal process is now defined as a process that transmits a nonzero amount of a conserved quantity at each moment in its history. Conserved quantities are quantities so characterized in physics: linear momentum, angular momentum, charge, and so on. A causal interaction is an intersection of world

lines associated with causal processes involving exchange of a conserved quantity. Finally, a process transmits a conserved quantity from  $A$  to  $B$  if it possesses that quantity at every stage without any interactions that involve an exchange of that quantity in the half-open interval  $(A, B]$ .

One may doubt that this new theory really avoids reliance on counterfactuals, but an even more fundamental difficulty is that it still does not adequately deal with the problem of causal or explanatory relevance described above. That is, we still face the problem that the feature that makes a process causal (transmission of some conserved quantity or other) tells us nothing about which features of the process are causally or explanatorily relevant to the outcome we want to explain. For example, a moving billiard ball will transmit many conserved quantities (linear momentum, angular momentum, charge, etc.), and many of these may be exchanged during a collision with another ball. We still face the problem of singling out the linear momentum of the balls, rather than these other conserved quantities, as the property that is causally relevant to their subsequent motion. In cases in which there appears to be no conservation laws governing the explanatorily relevant property (i.e., cases in which the explanatorily relevant variables are not conserved quantities), this difficulty seems even more acute. Properties like “having ingested birth control pills” and “being pregnant” do not themselves figure in conservation laws. One may say that both birth control pills and hexed salt are causal processes because both consist, at some underlying level, of processes that unambiguously involve the transmission of conserved quantities like momentum and charge, but this observation does not by itself tell us what, if anything, about these underlying processes is relevant to pregnancy or dissolution in water.

In a still more recent paper, Salmon (1997) conceded this point. He agreed that the notion of a causal process cannot by itself capture the notion of causal and explanatory relevance. He suggested, however, that this notion can be adequately captured by appealing to the notion of a causal process *and* information about statistical relevance relationships (i.e., information about conditional and unconditional (in)dependence relationships), with the latter capturing the element of counterfactual dependence that was missing from his previous account: “I would now say that (1) statistical relevance relations, in the absence of information about connecting causal processes, lack explanatory import and that (2) connecting causal processes, in the absence of statistical relevance relations, also lack explanatory import” (p. 476). In other words, the fact that whether Ms. Jones (gender =  $W$ ) ingests birth control pills ( $B$ ) is causally relevant to whether she becomes pregnant ( $P$ ) but not to the pregnancy of Mr. Jones (gender =  $-W$ ) is captured by statistical relevance relationships like the following:

$$\Pr(P/B \cdot W) \neq \Pr(P/-B \cdot W) \text{ but } \Pr(P/B \cdot -W) = \Pr(P/-B \cdot -W)$$

in *combination* with information about connecting causal processes.

This suggestion is not developed in any detail in Salmon’s paper, but I am skeptical that it can be made to work. We observed in previous chapters that

statistical relevance relationships often greatly underdetermine the causal relationships among a set of variables. What reason is there to suppose that appealing to the notion of a causal process, in Salmon's sense, will always or even usually remove this indeterminacy? We also noted that the notion of a causal process cannot capture fine-grained notions of relevance between properties, that there can be causal relevance between properties, instances of which (at least at the level of description at which they are characterized) are not linked by spatiotemporally continuous processes or transference of conserved quantities and that properties can be so linked without being causally relevant (recall the chalk mark that is transmitted from one billiard ball to another). As long as it is possible (and why should it not be?) for different causal claims to imply the same facts about statistical relevance relationships and for these claims to differ in ways that cannot be fully cashed out in terms of Salmon's notions of causal processes and interactions, this new proposal will fail as well.

## 8.5 Unificationism

The basic idea of the *unificationist* account is that explanation is a matter of providing a unified account of a range of different phenomena. This idea is unquestionably intuitively appealing. Theories that unify a range of different phenomena, previously dealt with by distinct theories, are in an obvious sense more general than these previous theories, and it is plausible that generality is at least sometimes an explanatory virtue. Moreover, theory unification has clearly played an important role in science; paradigmatic examples include Newton's unification of terrestrial and celestial theories of motion and Maxwell's unification of electricity and magnetism. The key question, however, is whether our intuitive notion (or notions) of unification can be made more precise in a way that allows us to recover the features that we think good explanations should possess.

Michael Friedman (1974) is an important early attempt to do this. Friedman's formulation of the unificationist idea was subsequently shown to suffer from various technical problems, and subsequent development of the unificationist treatment of explanation has been most associated closely with Philip Kitcher (especially 1989). My discussion focuses on Kitcher's views.

To explain Kitcher's framework, we need to introduce some of his technical vocabulary. A *schematic sentence* is a sentence in which some of the nonlogical vocabulary has been replaced by dummy letters. To use Kitcher's examples, the sentence "Organisms homozygous for the sickling allele develop sickle cell anemia" is associated with a number of schematic sentences, including "Organisms homozygous for  $A$  develop  $P$ " and "For all  $X$  if  $X$  is  $O$  and  $A$  then  $X$  is  $P$ ." *Filling instructions* are directions that specify how to fill in the dummy letters in schematic sentences. For example, filling instructions might tell us to replace  $A$  with the name of an allele and  $P$  with the name of a phenotypic trait in the first of the above schematic sentences. *Schematic arguments* are sequences

of schematic sentences. *Classifications* describe which sentences in schematic arguments are premises and conclusions and what rules of inference are used. An *argument pattern* is an ordered triple consisting of a schematic argument, a set of sets of filling instructions, one for each term in the sentence of the schematic argument, and a classification of the schematic argument. The more restrictions an argument pattern imposes on the arguments that instantiate it, the more *stringent* it is said to be.

Roughly speaking, Kitcher's guiding idea is that explanation is a matter of deriving descriptions of many different phenomena by using as few and as stringent argument patterns as possible over and over again: the fewer the patterns used, the more stringent they are, and the greater the range of different conclusions derived, the more unified our explanations. Kitcher summarizes this view as follows: "Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same the same pattern of derivation again and again, and in demonstrating this, it teaches us how to reduce the number of facts we have to accept as ultimate" (1989, p. 423).

Kitcher does not propose a completely general theory of how the various considerations he describes—number of conclusions, number of patterns and stringency of patterns—are to be traded off against one another, but he does suggest that it often will be clear enough what these considerations imply about the evaluation of particular candidate explanations. His basic strategy is to attempt to show that the derivations we regard as good or acceptable explanations are instances of patterns that, taken together, score better according to the criteria just described than the patterns instantiated by the derivations we regard as defective explanations. Following Kitcher, let us define the *explanatory store*  $E(K)$  as the set of argument patterns that maximally unifies  $K$ , the set of beliefs accepted at a particular time in science. Showing that a particular derivation is a good or acceptable explanation is then a matter of showing that it belongs to the explanatory store.

As an illustration, consider Kitcher's treatment of the problem of explanatory asymmetries. Our present explanatory practices, call these  $P$ , are committed to the idea that derivations of a flagpole's height from the length of its shadow are not explanatory. Kitcher compares  $P$  with an alternative systemization in which such derivations are regarded as explanatory. According to Kitcher,  $P$  includes the use of a single "origin and development" (OD) pattern of explanation, according to which the dimensions of objects (artifacts, mountains, stars, organisms, etc.) are traced to "the conditions under which the object originated and the modifications it has subsequently undergone" (1989, p. 485). Now consider the consequences of adding to  $P$  an additional pattern  $S$  (the shadow pattern), which permits the derivation of the dimensions of objects from facts about their shadows. Because the OD pattern already permits the derivation of all facts about the dimensions of objects, the addition of the shadow pattern  $S$  to  $P$  will increase the number of argument patterns in  $P$  and will not allow us to derive any new conclusions. On the other hand, if we were to drop OD from  $P$  and replace it with the shadow pattern, we would have no net change in the number of patterns in  $P$ , but would be able to derive far fewer



conclusions than we would with *OD*, because many objects do not have shadows (or enough shadows) from which to derive all of their dimensions. Thus, *OD* belongs to the explanatory store, and the shadow pattern does not.

Kitcher's treatment of other familiar problem cases is similar. For example, he notes that we believe that an explanation of why some sample of salt dissolves in water that appeals to fact that the salt is hexed and the generalization (*H*) that all hexed salt dissolves in water is defective, at least in comparison with the standard explanation that appeals just to the generalization that (*D*) all salt dissolves in water. He suggests that the "basis for this belief" is that the derivation that appeals to (*H*) instantiates an argument pattern that belongs to a totality of patterns that is less unifying than the totality containing the derivation that appeals to (*D*). In particular, an explanatory store containing (*H*) but not (*D*) will have a more restricted consequence set than a store containing (*D*) but not (*H*), because the latter but not the former allows for the derivation of facts about the dissolving of unhexed salt in water. And the addition of (*H*) to an explanatory store containing (*D*) will increase the number of patterns without any compensating gain in what can be derived.

Kitcher acknowledges that there is nothing in the unificationist account per se that requires that all explanation be deductive: "There is no bar in principle to the use of non-deductive arguments in the systemization of our beliefs." Nonetheless, "the task of comparing the unifying power of different systemizations looks even more formidable if nondeductive arguments are considered," and in part for this reason Kitcher endorses the view that "*in a certain sense, all explanation is deductive*" (1989, p. 448). We will examine how he proposes to deal with ostensibly nondeductive explanations below.

What is the role of causation in this account? Kitcher claims that "the 'because' of causation is always derivative from the 'because' of explanation" (1989, p. 477). That is, our causal judgments simply reflect the explanatory relationships that fall out of our (or our intellectual ancestors') attempts to construct unified theories of nature. There is no independent causal order over and above this that our explanations must capture. Kitcher takes very seriously, even if in the end he perhaps does not fully endorse, standard empiricist or Humean worries about the epistemic accessibility and intelligibility of causal claims. Taking causal, counterfactual, or other notions belonging to the same family as primitive in the theory of explanation is problematic. Kitcher believes it is a virtue of his theory that it does not do this. Instead, he proposes to begin with the notion of explanatory unification, characterized in terms of constraints on deductive systemizations, where these constraints can be specified in a quite general way that is independent of causal or counterfactual notions, and then show how the causal claims we accept derive from our efforts at unification.

## 8.6 Unification and Causation

Kitcher's account is thus fundamentally different in motivation from the manipulationist account. Unlike Kitcher, I see no reason to suppose that all

explanation is at bottom deductive. Moreover, my view is that when causal explanations are deductive, we should see their deductive structure as having explanatory import only to the extent that it traces or represents an independently existing causal order or set of dependency relationships, as revealed in facts about the outcomes of hypothetical experiments. Whereas Kitcher sees our judgment that the height of a flagpole causes or explains the length of its shadow as deriving from the fact that the unification achieved by the deductive systemization associated with our present explanatory practices *P* (which countenance derivations running from the height to the length as explanatory but not vice versa) is superior to alternative systemizations (which countenance shadow to height derivations), the manipulationist account traces this asymmetry to the fact that there are interventions on the flagpole's height that will change the length of its shadow, but no intervention on the length of the shadow that will change the height. On the manipulationist account, these facts about the outcomes of hypothetical experiments are primary, and the derivation running from the flagpole's height to the shadow length is explanatory because it mirrors these facts. Similarly, on my account, the basis of our judgment about the explanatory irrelevance of hexing salt has to do with the fact that manipulating whether the salt is hexed will make no difference to whether it dissolves, rather than deriving from the comparison between alternative deductive systemizations that Kitcher describes.

The idea that the causal order, as reflected in facts about the outcomes of hypothetical experiments, is independent of and prior to our efforts to represent it in deductive schemes (or in any other way) is central to my account. As argued in chapters 5 and 7, valid derivations and other sorts of mathematical manipulations do not automatically successfully trace or represent causal relationships. We can often derive the occurrence of a cause from the occurrence of its effect just as readily as we can run the derivation in the opposite direction. Moreover, on my view, there is no general reason to suppose that the latter derivation will belong to a deductive systemization that is more unified than the former: considerations having to do with unification do not automatically pick out those derivations that are explanatory from those that are not.

As an illustration of this general point, consider Kitcher's account of explanatory asymmetries. Kitcher's treatment of the flagpole example depends heavily on the contingent truth that some objects do not cast enough shadows to recover all of their dimensions. But it seems to be part not just of common sense, but of currently accepted physical theory that it would be inappropriate to appeal to facts about the shadows cast by objects to explain their dimensions even in a world in which all objects cast enough shadows that all their dimensions could be recovered. It is unclear how Kitcher's account can recover this judgment.

The matter becomes clearer if we turn our attention to a variant example in which, unlike the shadow example, there are clearly just as many backwards derivations from effects to causes as there are derivations from causes to effects. Consider, following Barnes (1992), a time-symmetric theory like Newtonian mechanics, applied to a closed system like the solar system. Call

derivations of the state of motion of planets at some future time  $t$  from information about their present positions (at time  $t_0$ ), masses, velocities, the forces incident on them at  $t_0$ , and the laws of mechanics *predictive*. Now contrast such derivations with *retrodictive* derivations in which the present motions of the planets are derived from information about their future velocities and positions at  $t$ , the forces operative at  $t$ , and so on. It looks as though there will be just as many retrodictive derivations as predictive derivations, and each will require premises of exactly the same general sort (information about positions, velocities, masses, etc.) and the same laws. Thus, the pattern or patterns instantiated by the retrodictive derivations look(s) exactly as unified as the pattern or patterns associated with the predictive derivations. However, we ordinarily think of the predictive derivations and not the retrodictive derivations as explanatory and the present state of the planets as the cause of their future state and not vice versa. It is again far from obvious how considerations having to do with unification could generate such an explanatory asymmetry.

One possible response to this second example is to bite the bullet and argue that, from the point of view of fundamental physics, there really is no difference in the explanatory import of the retrodictive and predictive derivations, and that it is a virtue, not a defect, in the unificationist approach that it reproduces this judgment. Whatever might be said in favor of this response, it is not Kitcher's. His claim is that our ordinary judgments about causal asymmetries can be derived from the unificationist account. The example just described casts doubt on this claim. More generally, it casts doubt on Kitcher's contention that one can begin with the notion of explanatory unification, understood in a way that does not presuppose causal notions, and use it to derive the content of causal judgments.

## 8.7 The Heterogeneity of Unification

This conclusion is reinforced by a more general consideration: unification, as it figures in science, is a quite heterogeneous notion, covering many different sorts of achievements. Some kinds of unification consist in the creation of a common classificatory scheme or descriptive vocabulary where no satisfactory scheme previously existed, as when early investigators like Linnaeus constructed comprehensive and principled systems of biological classification. Another kind of unification involves the creation of a common mathematical framework or formalism that can be applied to many different sorts of phenomena, as when the systems of equations devised by Lagrange and Hamilton were first developed in connection with mechanics and then applied to domains like electromagnetism and thermodynamics. Still other cases involve what might be described as genuine physical unification, where phenomena previously regarded as having quite different causes or explanations are shown to be the result of a common set of mechanisms or causal relationships. Newton's demonstration that the orbits of the planets and the behavior of

terrestrial objects falling freely near the surface of the earth are due to the same force of gravity and conform to the same laws of motion was a physical unification in this sense.

Of these three kinds of activities only the third, physical unification, seems to have much intuitively to do with causal explanation. In particular, the kind of unification associated with adoption of a classificatory scheme often tells us little about causal relationships. Moreover, as historical studies have made clear, a similar point holds for formal or mathematical unification: the fact that we can construct a common mathematical framework for dealing with a range of different phenomena does not by any means automatically ensure that we have identified some set of common causal factors responsible for those phenomena (i.e., that we have produced a unified physical explanation of them).<sup>2</sup> For example, the mere fact that we can describe both the behavior of a system of gravitating masses and the operation of an electric circuit by means of Lagrange's equations does not mean that we have achieved a common explanation of the behavior of both or that we have "unified" gravitation and electricity in any physically interesting sense.

Just as there are different varieties of unification, these may be pursued for a range of different motives, only some of which seem to have much to do with a desire for explanation. For example, one reason scientists may seek to extend a mathematical framework from one domain to another is that they are familiar with the formalism, know how to solve problems or calculate with it, and find it tractable. This is a perfectly reasonable motive for a kind of unification, but one that may have little to do with explanation. Similarly, scientists may adopt a classificatory scheme because it is easy to use, facilitates communication, and leads to a large measure of intersubjective agreement. Again, although such schemes may involve a kind of unification, we should not assume that they are adopted for their explanatory virtues.

These considerations raise the following question: Is Kitcher's account of unification sufficiently discriminating or nuanced to distinguish those unifications having to do with explanation from other sorts of unification? The worry is that it is not. The conception of unification underlying Kitcher's account seems to be at bottom one of descriptive economy or information compression: deriving as much from as few patterns of inference as possible. Many cases of classificatory and purely formal unification involving a common mathematical framework seem to fit this characterization. Consider schemes for biological classification and schemes for the classification of geological and astronomical objects like rocks and stars. If I know that individuals belong to a certain classificatory category (e.g., *Xs* are mammals or polar bears), I can use this information to derive a great many of their other properties (*Xs* have backbones and hearts, their young are born alive, etc.), and this is a pattern of inference that can be used repeatedly for many different sorts of *Xs*. But despite the willingness of some philosophers to regard such derivations as explanatory, previous chapters have argued that scientific practice is to regard such schemes as "merely descriptive" and as telling us little or nothing about the causes or mechanisms that explain why *Xs* have backbones or hearts. Because

there is no well-defined notion of intervention that consists in changing whether an organism is a mammal or a polar bear, such classificatory schemes do not convey information that is relevant to manipulation and control and do not figure in the answers to a range of *w-questions*.

Another illustration of the same general point is provided by the numerous statistical procedures (factor analysis, cluster analysis, multidimensional scaling techniques) that allow one to summarize or represent large bodies of statistical information in an economical, unified way and to derive more specific statistical facts from a much smaller set of assumptions by repeated use of the same pattern of argument. For example, knowing the “loading” of each of  $n$  intelligence tests on a single common factor  $g$ , one can derive a much larger number  $(n(n-1)/2)$  of conclusions about pairwise correlations among these tests. Again, however, it is doubtful that by itself this “unification” tells us anything about the causes of performance on these tests.

A parallel point arises for what I called mathematical unification. Suppose that Lagrange’s equations are used to derive predictions about both the behavior of a system of gravitating masses and an electrical circuit. It would appear that there is a straightforward sense in which this involves the use of the same general pattern of derivation over and over again and hence that this ought to count as an explanatory unification on Kitcher’s theory, at least in comparison with the use of two quite different sets of laws and derivational patterns for these different domains. However, this “unification” does not seem to involve a common set of causal or explanatory factors.

Of course, it is possible to respond that the argument patterns used in connection with these two domains are really quite different: they require that different Lagrangians be plugged into Lagrange’s equations and so on or, what I take to be the same thing, that they involve the “same” pattern only for a relatively nonstringent characterization of that pattern, and that when this is taken into account, the case does not turn out to be one of explanatory unification after all. This response, however, raises the question of whether the sorts of judgments that Kitcher’s account requires about number and stringency of patterns and number of conclusions and how these should be traded off against one another can really be made in a nonarbitrary way. Recall that in connection with the flagpole example, Kitcher speaks of a single *OD* pattern of explanation and invites us to compare our present situation, in which we rely just on this pattern to explain the dimensions of objects, with alternative systemizations in which we add the shadow pattern *S* to *OD* or replace *OD* with *S*. But, on the face of things, the argument pattern(s) we use to explain the dimensions of artifacts, which might appeal to the intentions of the artisan, seem(s) quite different from the argument patterns we use to explain the dimensions of mountains, which will appeal to various geological processes, and these are in turn quite different from the argument patterns we might use to explain the dimensions of living organisms. Obviously, a further splitting of argument patterns might be made in each of these categories. At the very least, these considerations show that the *OD* pattern is *very* nonstringent. What entitles us to frame the comparison of our present practices with *S* in terms of

the *OD* pattern rather than some more stringent alternative that could also be used to derive flagpole lengths?

In the present case, it is arguable, although by no means obviously true, that any reasonable alternative to *OD* will also yield the result that the direction of explanation runs from flagpole to shadow (cf. Barnes 1992). In other cases, however, the conclusions we reach about causal and explanatory relationships may be quite sensitive to exactly how we count or individuate patterns and our judgments of relative stringency. If reasonable people may disagree about how to count patterns or evaluate stringency, this seems to undermine the use of the unificationist framework to evaluate competing explanations.

The threat of arbitrariness or subjectivity in assessments of relative unification arises not just in comparisons of stringency and number of argument patterns but also in comparisons of the “number” of “different” descriptions of phenomena (or number of different conclusions regarding phenomena) that may be derived using a given argument pattern. As noted in connection with the MRL theory of law, one kind of comparison of the “size” of the sets of conclusions derivable from different theories seems unproblematic: if the consequence set of  $T_1$  is a proper subset of the consequence set of  $T_2$ , then there is an uncontroversial respect in which  $T_2$  permits the deduction of more conclusions than  $T_1$ . However, this yields at best a basis for ordinal comparisons: it provides no basis for assigning a cardinality to the “number” of conclusions that follow from  $T_1$  and  $T_2$ , a number that might then be traded off against number and stringency of patterns required to derive those conclusions in order to assess the degree of unification associated with  $T_1$  and  $T_2$ . Moreover, the resulting ordering is only partial: we have no basis for comparing consequence sets when one is not a proper subset of the other. Given two theories  $T_1$  and  $T_2$ , if the consequence set of  $T_1$  is a proper subset of the consequence set of  $T_2$  and if  $T_2$  makes use of the same number of argument patterns as  $T_1$ , and these patterns are of equal stringency or, if  $T_2$  uses fewer or more stringent patterns, then we may conclude that  $T_2$  is more unified than  $T_1$ . A similar conclusion will follow if  $T_1$  and  $T_2$  have the same consequence set and  $T_2$  uses fewer or more stringent argument patterns. However, we have as yet no basis for making other judgments of comparative unification. In particular, we have no basis for seemingly intuitive judgments like the following: when Newtonian mechanics and gravitational theory are used to derive a range of different conclusions involving both celestial and terrestrial motions, these phenomena are very different from each other, whereas if we use Newton’s laws to derive the acceleration of a one-gram test particle under a 1.0 and under a 1.1 Newton force, these phenomena are not interestingly different from one another. It is judgments of the former sort that seem to underlie our sense that Newtonian mechanics is a highly unifying theory.

In her important book-length study of unification, Margaret Morrison (2000) notes, in an observation that closely parallels the remarks about “number of different conclusions” in my discussion of the MRL theory of law in chapter 6, that the most natural account of the sense in which celestial and terrestrial

motions are “very different” from one another is that these phenomena are very different from the perspective of a prior theory, like Aristotle’s, that regards the causes of celestial and terrestrial motions as fundamentally distinct. From the perspective of other prior theories, such as Cartesian mechanics, which assume, as Newton did, that celestial and terrestrial motions have broadly the same causes, these phenomena do not seem fundamentally different. This suggests that our judgment of whether a range of phenomena are importantly different from one another, or about the cardinality of the set of “different” conclusions that may be derived from a theory, must be relativized to our prior epistemic situation in some way; it depends on what is previously known or believed about those phenomena.<sup>3</sup> This sort of relativization perhaps might be regarded as undisturbing if the only sorts of judgments about unification that we need to make are judgments of comparative unification (e.g., judgments to the effect that Newton’s theory is more unified than Aristotle’s). However, we also seem to need noncomparative judgments of unification, for reasons described below.

We may get some further purchase on this issue by contrasting the unificationist approach with the manipulationist account. Both approaches take generality to be an important desideratum in connection with explanations. However, they understand the notion of generality in very different ways. On the most natural interpretation of the unificationist account, generality is cashed out in terms of breadth of scope—in terms of the range of actual phenomena to which a theory applies. By contrast, on the manipulationist account, generality is cashed out in terms of range of invariance. Recall from chapter 6 that a generalization can have very wide scope while being invariant only under a narrow range of interventions, or indeed without being invariant under any interventions at all. Conversely, a generalization can have narrow scope while being invariant under a wide range of interventions. In my view, the explanatory depth of a generalization is connected to its range of invariance rather than its scope; hence, the unificationist approach focuses on the wrong sort of generality in explanations.

Two examples from chapter 6 illustrate the basic difficulty. The generalization (*K*) specifying, for each spatial region of the universe, that the microwave radiation background left over from the big bang has very wide scope. (*K*) is a unifying generalization that could be used, over and over again, in the derivation of many different phenomena. Nonetheless, it does not follow from this fact that (*K*) is invariant over a wide range of interventions. Indeed, because (*K*) is not naturally interpretable as a change-relating generalization, it is not clear that there are even any well-defined interventions with respect to (*K*). The presence of the microwave background in one spatial region certainly does not cause or explain its presence in any other region; the uniformity of the background in different regions is instead caused by the initial conditions obtaining in the very early universe. (*K*) describes an extremely pervasive uniformity, but pervasiveness has to do with scope, not invariance. On my account, (*K*) is not an explanatory generalization. It looks as though unificationist accounts reach the opposite conclusion.

As a second illustration, consider two different neural circuits  $N_1$  and  $N_2$ .  $N_1$  is highly conserved: it is found in many different kinds of organisms, as diverse as snails and human beings, and is involved in many different kinds of overt behavior, although the operation of  $N_1$  itself is governed by the same simple generalization in each case. (Think of a Hebbian learning rule.) By contrast,  $N_2$  is found only in a certain species of snail and is involved in only a single characteristic bit of behavior, and is also governed by a single learning rule. There is an obvious sense in which the generalization governing  $N_1$  has a much greater scope than the generalization governing  $N_2$ . If derivations of different sorts of overt behavior in different species that appeal to the generalization governing  $N_1$  count as derivations of significantly “different” phenomena—and it is unclear why they should not—the unificationist account seems to yield the conclusion that such derivations provide more unified and hence better or deeper explanations than derivations that appeal to the generalization governing  $N_2$ , again on the grounds that the generalization governing  $N_1$  has greater scope. It is not obvious that this is the correct assessment. Again, the invariance-based account avoids this conclusion, because it takes range of invariance rather than scope to be relevant to explanatory depth. Alternatively, if it is claimed instead that the behaviors produced by  $N_1$  in different species do not count as different phenomena, we are owed an account of the principles governing the counting of phenomena that yield this conclusion.

### 8.8 Another Difficulty: The Winner-Take-All Conception of Explanatory Unification

There is yet another fundamental difficulty with the unificationist account. On the one hand, it seems that any plausible version of that account must yield the conclusion that generalizations and theories can sometimes be explanatory with respect to some set of phenomena, even though more unifying explanations of those phenomena are known. For example, the regression equation described in chapter 7 relating water to plant height seems explanatory even though there are far deeper, more biologically grounded explanations of plant growth. Similarly, Galileo’s law can be used to explain facts about the behavior of falling bodies even though it furnishes a less unifying explanation than the laws of Newtonian mechanics and gravitational theory; the latter are in turn explanatory even though the explanations they provide are less unified than those provided by General Relativity; the theories of Coulomb and Ampere are explanatory even though the explanations they provide are less unified than the explanations provided by Maxwell’s theory; and so on. If we reject this idea, we seem led to the conclusion that in any domain, only the most unified theory that is known is explanatory at all; everything else is nonexplanatory. Call this the winner-take-all conception of explanatory unification.

The winner-take-all conception gives up on the apparently very natural idea, which one would think the unificationist would wish to endorse, that an



explanation can provide less unification than some alternative, and hence be less deep or less good, but still qualify as somewhat explanatory. Moreover, once one accepts the idea that theories that are less unifying than the most unifying *known* theory are not explanatory, it is hard to reject the further conclusion that the only theories that are explanatory are the most unifying theories that will ever be discovered (or perhaps the most unifying theories that exist, whether or not they will ever be known). To say that  $T_1$  is explanatory now, in virtue of the unification it achieves, but becomes unexplanatory once some even more unifying theory  $T_2$  becomes known, is to relativize explanatory success to our knowledge situation in a very radical (and, one would think, undesirable) way. But the conclusion that  $T_1$  is unexplanatory now, no matter how much unification it achieves, as long as some other more unifying theory exists or will one day be discovered, also seems unsatisfactory.

Kitcher's treatment of the problems of explanatory irrelevance and explanatory asymmetry seems to require the winner-take-all assumption I have been criticizing. Why is it that we cannot appeal to the fact that this particular sample of salt has been hexed to explain why it dissolves? According to Kitcher, any explanatory store of which this "explanation" is a part will be "less unified" than a competing explanatory store according to which the dissolving of the salt is explained by appeal to the generalization that all salt dissolves in water. Similarly, the reason we cannot explain the height of a flagpole in terms of the length of its shadow is that explanations of lengths of objects in terms of facts about shadows do not belong to the "set of explanations" that "collectively provides the best systemization of our beliefs" (1989, p. 430). This analysis clearly requires the winner-take-all idea that an explanation  $T_1$  that is less satisfactory from the point of view of unification than some competing alternative  $T_2$  is *unexplanatory*, rather than merely *less explanatory*, than  $T_2$ . If Kitcher were to reject the winner-take-all idea and hold instead that even if  $T_2$  is more unified than  $T_1$ , it does not automatically follow that  $T_1$  is unexplanatory, then his solution to the problems of explanatory irrelevance and asymmetry would no longer be available: his conclusion should be that an "explanation" of Mr. Jones's failure to get pregnant in terms of his ingestion of birth control pills is genuinely explanatory, although less so than the alternative explanation that invokes his sex, and similarly for a derivation of the height of a flagpole from the length of its shadow.

Intuitively, the problem is that we need a theory that captures several different possibilities. On the one hand, there are generalizations and associated putative explanations (such as the generalization relating barometric pressure to the occurrence of storms and the generalization relating the hexing of salt to its dissolution in water) that are not explanatory at all; they fall below the threshold of explanatoriness. On the other hand, above this threshold there is something more like a continuum: a generalization can be explanatory but provide less deep or good explanations than some alternative. What we have just seen is that the unificationist account has difficulty simultaneously capturing both of these possibilities: either there is no threshold (every derivation is explanatory to some extent; it is just that some derivations belong to

systemizations that are less unifying and hence less explanatory than others), or else there is no continuum (only the most unifying systemizations are explanatory). By contrast, the account I have proposed captures the idea that there is both a threshold and a continuum in a very natural way. Some generalizations are not invariant under any (testing) interventions at all and hence are nonexplanatory. Other generalizations are invariant under some testing interventions (and answer some *w-questions*) and hence are above the threshold of explanatoriness, although they are less invariant and answer a narrower range of *w-questions* than others and hence are less explanatory.

## 8.9 The Epistemology of Unification

I observed above that, according to Kitcher, causal knowledge derives from our efforts at unification. However, as Kitcher also recognizes, it is highly implausible that most individuals deliberately and self-consciously go through the process of comparing competing deductive systemizations with respect to number and stringency of patterns and number of conclusions in order to determine which is most unifying. His response to this observation is to hold that most people acquire causal knowledge by absorbing the “lore” of their community, where this lore does reflect previous systematic efforts at unification. He writes that “our everyday causal knowledge is based on our early absorption of the theoretical picture of the world bequeathed to us by our scientific tradition” (1989, p. 469).

It is not easy to see how this suggestion is supposed to work. First, although it is surely true that individual human beings acquire a substantial amount of causal knowledge by cultural transmission, it is also obvious that not all causal knowledge is acquired in this way. Some causal knowledge individuals acquire involves learning from experience. Moreover, unless we are willing to make extremely implausible assumptions about the innateness of a large number of specific causal beliefs, the stock of socially transmitted causal knowledge must itself have been initially acquired in a way in which learning from experience played an important role. The question that then arises is how this process of learning from experience is supposed to work on a view like Kitcher’s about the source of our causal knowledge. If, as Kitcher claims, “the idea that any one individual justifies the causal judgments that he/she makes by recognizing the patterns of argument that best unify his/her beliefs is clearly absurd” (1989, p. 436), just what is it that is going on at the individual level when people learn from experience? One possibility is that, although individuals do not knowingly go through the process of comparing the degree of unification achieved by alternative systemizations when they acquire new causal knowledge by learning from experience, they go through this process tacitly or unconsciously, perhaps because of some general disposition of the mind to seek unification. However, Kitcher does not seem to endorse this idea and it does not seem to fit very well with his emphasis on the social transmission of causal information. Moreover, it looks as though even

unconscious unification requires very sophisticated cognitive abilities (construction and comparison of different deductive systemizations, etc.) that it is implausible to attribute to many causal learners, such as small children.

One natural interpretation of the passages quoted above and others in Kitcher (1989) is this: a social process of comparing alternative systemizations of beliefs and drawing out their deductive consequences occurs at the community level, with groups of people making arguments to one another about which overall deductive systemizations best unify the beliefs of the community as a whole. Particular causal beliefs are justified at the community level by being shown to be part of the best overall systemization of the beliefs of the community, and are then passed on from the common community stock to individuals via a process of social transmission.

An obvious problem with this picture is that the communitywide process of justification must be carried out in some fashion by individual actors. If, as appears to be the case, there are many societies in which no one possesses an explicit or clearly articulated concept of a deductively valid argument or is very skilled at drawing out the deductive consequences of beliefs or possesses explicit versions of Kitcher's concepts of number and stringency of argument patterns, how exactly are community beliefs that reflect the operation of these notions supposed to form? If, as Kitcher concedes, it is psychologically unrealistic to assume that individual human beings deliberately and self-consciously go through the process of comparing alternative systemizations when they acquire causal beliefs through experience, why is it any more realistic to suppose that this process somehow occurs through the interactions of individual actors at the community level?

There is a second, related difficulty. Assume, for the sake of argument, that it is desirable to have a unified belief system in Kitcher's sense—whether because unification is connected to explanation and the latter is intrinsically valuable or because unification is connected to other goals that are desirable. It is still not obvious why it would be valuable to have a set of beliefs that are a smallish proper subset of the beliefs that compose such a unified system, which is what most people seem to have, given Kitcher's views about the transmission of causal knowledge. Recall Kitcher's basic picture: when I acquire the belief that, say, whether salt is hexed is causally irrelevant to whether it dissolves and that whether it is placed in water is causally relevant, I acquire a fragment of the community's overall systemization *S*. But adding a fragment of *S* or even a number of fragments of *S* to my belief store may not result in *my* having a belief system that is unified or that facilitates whatever epistemic goals are associated with unification. Of course, if I end up adding all or most of *S* to my belief store, I will have at that point (although not before) a set of beliefs that is unified and that brings with it all of the benefits of unification. But I take it that Kitcher agrees that it is often completely unrealistic to suppose that most people possess the full systemization *S* that best unifies all of the beliefs in their community. This seems to be true, for example, of our own epistemic community, in which knowledge, especially scientific knowledge, is highly dispersed among a small group of experts and in which no single person's mind (and still less the typical

member's mind) contains or operates in accordance with the systemization that best unifies the beliefs of the entire community. More generally, it seems unlikely that the different portions  $B_i$  of the community systemization  $S$  that various individuals  $i$  acquire by means of cultural transmission will be in each case highly unified systemizations. In short, it is a major problem with the cultural transmission story that it is hard to see how unification could be cognitively or practically valuable unless it characterizes the belief systems of individuals and not just the community. However, taking the sort of unification Kitcher describes to characterize individual belief systems seems *prima facie* psychologically unrealistic. I do not mean to claim that there is no way of making sense of the acquisition of causal knowledge on the unificationist picture, but I do think that a great deal more needs to be said about how this works.

### 8.10 Deductive Chauvinism

Kitcher is a self-described "deductive chauvinist": he holds that all explanation is deductive. His basic strategy in defending this claim is to argue that purported explanations that appear to be nondeductive can, to the extent that they are genuinely explanatory, be reconstrued as deductive arguments. In some cases, this strategy seems quite plausible. For example, Kitcher argues, as I have elsewhere (Woodward 1989), that many quantum mechanical explanations are deductive. In cases in which the state vector is not in an eigenstate of some observable, there is no compelling reason to interpret quantum mechanics as explaining why that observable takes the value it does on measurement. Instead, it is more natural to interpret the theory as explaining facts about the probability or expectation value of such outcomes. Such explananda can be deduced from the fundamental assumptions of the theory and information about initial conditions.

In other cases, however, this strategy of reconstruing purported nondeductive explanations as deductive seems more problematic. Consider Scriven's (1959a) example of the mayor with paresis. It is natural to think of the mayor's condition as explained by his untreated syphilis, even though not all syphilitics develop paresis and there is no deductive argument from the premise that he has syphilis and other known generalizations about the relationship between syphilis and paresis to the conclusion that he has paresis. Kitcher's response to this example is to suggest that what is explained is not why the mayor developed syphilis while other syphilitics did not, but rather a different explanandum, one that is deducible from known premises. He writes:

We cannot explain why the mayor, rather than other syphilitics, contracted paresis. However, the statement that the mayor had syphilis may answer a different why-question. Suppose that the why-question has an explicit presupposition: "Given that one of the townspeople contracted paresis, why was it the mayor?" Only syphilitics get paresis and he was

the only syphilitic in town. Note in this case we can deduce the *explanandum* from the presupposition of the question and the information given in the answer. Under these circumstances we can vindicate the idea that the statement that the mayor had syphilis is part of an explanation of something like his getting paresis—but the explanation is deductive. (1989, p. 457)

He adds that this is an instance of a more general explanatory strategy: “Sometimes we show that a system is in state  $X$  by presupposing that it is in one of the states  $\{X, Y_1, \dots, Y_n\}$  and demonstrating that it cannot be in any of the  $Y_i$ ” (p. 457).

The problem with this suggestion is that it is too permissive: it makes it far too easy to explain. Let us vary Kitcher’s example somewhat: suppose there are  $n + 1$  syphilitics in town  $\{X, Y_1 \dots Y_n\}$ , only one of whom, the mayor =  $X$ , develops paresis. Can we use this information to explain why (i) the mayor developed paresis, on the grounds that given the explicit presuppositions that (ii) at least one person did, and (iii) none of  $Y_1 \dots Y_n$  did, and the additional premises that (iv) the only syphilitics in town are  $\{X, Y_1 \dots Y_n\}$ , and (v) only syphilitics develop paresis, it is deducible that (i) the mayor did? This version of the example differs from Kitcher’s in that there are now  $n + 1$  syphilitics rather than just one, but preserves the deducibility of (i) from the presuppositions (ii)–(iii) and other background information (iv)–(v). Although (ii)–(v) do provide a reason to believe that (i) is true, it is dubious that they provide a deductive *explanation* of (i). As argued throughout this book, an explanation of an outcome must cite factors on which that outcome depends and possibly generalizations describing dependency relations in which the outcome figures. The explanation should not cite factors that are irrelevant to the outcome. Clearly, the mayor’s paresis depends only on *his* syphilis (and perhaps on other constitutional facts about him); it does not depend on the facts described by (ii)–(iv), which have to do with whether others have syphilis or paresis. A similar assessment applies to Kitcher’s original example: it is no part of the explanation of why the mayor developed paresis that no one else in town has syphilis. (The explanation for the mayor’s paresis would be the same, regardless of whether there are other syphilitics in town.) More generally, Kitcher’s treatment collapses the distinction between explaining why an outcome occurs and providing a reason for thinking that the outcome has occurred. As repeatedly emphasized, premises that figure in a deductive argument that some outcome has occurred may fail to be part of an explanation of why that outcome has occurred and derivational structure does not automatically mirror explanatory or dependency relationships.

As explained in chapter 5, my preferred treatment of examples like those under discussion is quite different from Kitcher’s. Rather than looking for a reconstrual of

(8.10.1) The mayor’s latent syphilis caused his paresis

that shows this to be a deductive explanation, my view is that we should see (8.10.1) as explanatory in virtue of its answering a *w-question* or exhibiting

a condition on which the mayor's paresis depends. (8.10.1) conveys the information that if the mayor had not contracted syphilis, he would not have developed paresis and that, given that he does have syphilis, the probability of paresis is greater than the 0 value it would have in the absence of syphilis. (8.10.1) is explanatory in virtue of conveying this information even though it does not provide information from which we can deduce the occurrence of the mayor's paresis. (8.10.1) explains the contrast between the mayor's developing paresis and his not developing paresis, although (as Kitcher says) it does not explain why the mayor rather than other syphilitics developed paresis. On this view of the matter, in contrast to Kitcher's, the factors that figure in the explanation of the mayor's syphilis are just those on which his syphilis depends. We thus see an additional respect in which Kitcher's unificationist conception of explanation and the idea that explanation involves capturing dependency relationships lead to a quite different treatment of particular examples.

### 8.11 Concluding Assessment

I think it follows from these critical observations that Kitcher's particular way of cashing out the idea of explanatory unification needs to be rethought. However, it certainly does not follow that the more general idea that there is a deep connection between explanation and (some appropriately characterized) notion of unification is misguided. Indeed, as already intimated, there is at least one important respect in which such a connection will hold within the interventionist framework that I favor. This framework ties the depth of the explanations in which a generalization figures to the range of invariance of that generalization and the range of what-if-things-had-been-different questions that it can be used to answer, and generalizations and theories that score well along this dimension will (at least often) be relatively general and unifying. Thus, at least some of the intuitions and judgments about particular cases that motivate the unificationist model can be recovered in the interventionist approach. Whether the interventionist account can do full justice to all that is right about the connection between explanation and unification is a complex question I hope to revisit at a later time. My intention in this chapter has been merely to exhibit some differences between the manipulationist account and current versions of the unificationist and *CM* models and to suggest that there are reasons to prefer the former.

## Afterword

This book has defended a manipulationist account of explanation and causation: causal and explanatory relationships are relationships that are potentially exploitable for purposes of manipulation and control. The manipulationist theory is a (species of a) counterfactual theory of explanation and causation, but differs in a number of ways from the counterfactual theories presently in the literature. I have tried to show that this theory both captures a number of important features of scientific and ordinary practice in connection with causation and explanation and allows us to avoid the counterexamples and difficulties that infect alternative approaches, from the deductive-nomological model onward. I have also argued that it provides a basis for criticism of a number of features of explanatory practice, particularly in the social sciences.

The sense of “potentially exploitable” to which I have appealed is an idealized one: what matters is not whether human beings can actually carry out manipulations on the magnitudes  $X$  and  $Y$  that figure in some candidate causal relationship, but rather whether the relationship correctly describes how  $Y$  would change if a change in  $X$  were produced by a special sort of causal process that I call an *intervention*. The notion of an intervention is an abstract representation of a human experimental manipulation that is stripped of its anthropocentric elements and characterized in terms that make no reference to human beings or their activities. Relationships that correctly tell us how  $Y$  would change under an intervention on  $X$  are *invariant*. The notion of an invariant relationship, I argue, is more helpful than the notion of a law of nature (the notion on which philosophers have traditionally tended to rely) for understanding explanation and causation.

Although this is a long book, I have left many issues unexplored. Aside from assuming that causal knowledge and explanations are cognitively valuable and worth acquiring, I have said nothing more specific about their place in inquiry. For example, I have not discussed the status of so-called inference to the best explanation: Is the fact that a hypothesis would, if true, provide an explanation of some known explanandum in itself a reason for believing that hypothesis? Or is it instead only a reason for pursuing or exploring that hypothesis, with reasons for belief requiring additional, independent evidence? Different accounts of explanation seem to fit more naturally with either the first or second of these alternatives, with (I am inclined to believe) the manipulationist account favoring the second.

I'm also very conscious of having failed to do justice to the full range of unificationist approaches to explanation. One thread in explanatory unification—*prima facie* rather different from the feature discussed in chapter 8—seems to involve the reduction of contingency or the ruling out of possibilities. Examples include cases in which a theory that accounts for certain explananda by making special assumptions about parameter values is replaced by a theory that avoids such special assumptions, perhaps by dispensing with the parameter altogether or by showing that the explananda would follow for all or almost all parameter values, and cases in which an explanandum is shown to be a mathematical truth by making certain identifications (Glymour 1980). By contrast, the manipulationist account involves the apparently different idea that explanation is a matter of showing how some possibilities depend on others; it involves tracing some contingencies to others, but needn't involve anything that looks much like a reduction in contingency.<sup>1</sup> What more can be said about the relationship between these two ideas and about their relative attractiveness as reconstructions of intuitive conceptions of explanation?

I also have not had space to explore the implications of the ideas developed here for areas of philosophy that are consumers of ideas about causation and explanation produced elsewhere. Even a very casual reader of work in philosophy of psychology, for example, will be struck by the extent to which current discussions of everything from reduction to mental content are still hostage to *DN*-inspired ideas. It is natural to wonder how these problems would look in other frameworks for thinking about cause and explanation.



## Notes

### Chapter 1 Introduction and Preview

1. As Mellor (1995) puts it, “Don’t just ask for the use, ask for the *point* of the use” (p. 230).
2. Obviously, this whole topic requires much more space than I can give it here. One obstacle to discussion, from my point of view, is that philosophers of language and psychologists have been very successful in appropriating concept talk for purely descriptive purposes, that is, the notion “concept of X” is now taken to be a notion that figures in semantic theory or a theory of mental representation. The more normative sorts of concerns that are important to my project play little or no role in such theories. I will also add that these normative concerns are strongly evident in the early literature on explanation, for example, Hempel (1965a).
3. The requirement that an invariant generalization must be stable under some interventions is a relatively weak, although far from vacuous, requirement. Various ways of strengthening this requirement without requiring invariance under all possible interventions are considered in chapter 6.
4. A similar list of desiderata is offered in Hughes (1993).

### Chapter 2 Causation and Manipulation

1. I have received helpful comments from many colleagues on the material in this chapter, but I owe particular thanks to two. Chris Hitchcock repeatedly saved me from major blunders. Clark Glymour, though thinking that it ignores most of the interesting issues about causation, nonetheless very generously sent me extremely detailed and searching comments.
2. For reasons described in Cowie (1999) and Elman, Bates, Johnson, Karmiloff-Smith, Parisi, and Plunkett (1996).
3. It is true that both Lewis and Salmon acknowledge the role of causal information in decision making. My point is that their theories of causation leave unclear why causal information is relevant to decision making.
4. A natural thought that will occur to many readers is that the most promising way of developing a probabilistic version of the regularity theory is to appeal to

“screening off” or conditional independence relationships. However, as will become apparent below, often a number of different causal relationships will be compatible with the same conditional independence relationships, showing that we cannot completely capture the content of the former in terms of the latter.

5. The contrast between the sort of correlational information that merely provides evidence that some outcome is more likely to occur (as when the information that a randomly selected individual has purchased TIAA insurance provides evidence regarding his longevity) and causal information that is relevant to bringing about that outcome is a central theme in discussions of evidential versus causal decision theory. See, for example, Gibbard and Harper (1976).
6. Spirtes et al. ([1993] 2000, chap. 9) contains a systematic comparison of the informativeness of experimentation and observation, assuming the Causal Markov and Faithfulness conditions (see below), in various sorts of real-life situations in which our abilities to measure or manipulate relevant variables are limited. Depending on the structure of the situation, observation may or may not be superior to experiment. By contrast, I abstract away from such issues, assuming an “in principle” framework in which all variables may be manipulated or measured.
7. For ease of exposition, I often use “variable” to describe both the properties, magnitudes, and so on related by causal claims and the representations we use to describe such properties. I believe that the resulting conflation of use and mention is (in this context) harmless.
8. This characterization is closely related to the characterization of “direct effect” in Pearl (2000a, p. 127).
9. See Spirtes et al. ([1993] 2000) and Pearl (2000a) for a more detailed discussion of the use of directed graphs to represent causal relationships.
10. Cartwright (2002) appeals to such an example to criticize claims about the connection between causation and manipulation in Hausman and Woodward (1999).
11. It may be tempting to suppose that the difficulty of explaining what “possible” in (i) means can be avoided by simply dropping (i) and formulating a manipulability theory just in terms of (ii), that is, by opting for a conditional rather than a conjunctive formulation of the theory. (This possibility is sympathetically explored, but not fully endorsed, by Ernest Sosa and Michael Tooley 1993, introduction.) This would be a mistake. To avoid trivialization, the conditional in (ii) cannot be understood as a material or strict conditional but must instead be understood as a counterfactual of some kind. An illuminating version of the manipulability theory thus needs to make clear how such counterfactuals are to be understood and what their truth conditions are. In particular, we need to know just what sort of possibility we should be envisioning when we envision the antecedent of a conditional along the lines of (ii). This in turn means that we cannot duck the question of what (i) means.
12. This distinction between intervening and conditioning is made very clearly in Meek and Glymour (1994) and in Pearl (2000a).
13. For additional discussion of autonomy, see Woodward (1995).
14. Within a probabilistic framework in which, following Pearl, *Set X* represents an intervention on *X*, modularity corresponds to the following requirement:  $Pr(X/Parents(X) \cdot Set Y) = Pr(X/Parents(X))$  for all *Y* distinct from *X*. See chapter 7

and Hausman and Woodward (1999) for further discussion and Cartwright (2001) for criticism.

15. To see this, substitute (3.5) into (3.4), obtaining  $Y$  as a function of  $X$  alone:  $Y = (a + bc)X$ . If  $a + bc = 0$ , changing  $X$  will have no overall effect on  $Y$ .
16. My colleague Chris Hitchcock (2001b) also distinguishes between these two notions of cause, describing the first as having to do with “net” effect and the second as having to do with “component effect along a causal route.”
17. If  $B$  and  $P$  interact with respect to  $T$ , then  $B$  will have a different impact on  $T$  among pregnant and nonpregnant women. However, unless  $B$  changes the probability of  $T$  in (at least) one of these populations,  $B$  will not count as a cause of  $T$ , and unless it raises the probability of  $T$ , it will not count as a positive cause of  $T$ , as figure 2.3.4 claims.
18. Note that in employing this reasoning we are assuming that intervening to fix the value of  $Z$  or of  $P$  does not alter the relationship between  $X$  and  $Y$  or between  $B$  and  $T$ ; that is, we are assuming that modularity holds.
19. Judea Pearl (2000a, p. 127) uses a closely related strategy to define a notion of “direct effect.” He writes, “The direct effect of  $X$  on  $Y$  is given by  $P(y/\text{set } x, \text{set } S_{xy})$  where  $S_{xy}$  is the set of all endogenous variables except  $X$  and  $Y$  in the system.” My discussion is indebted to his, but differs in several respects. To begin with, the restriction to endogenous variables doesn’t seem right because in a structure like that in figure 2.3.1, we need to set the exogenous variable  $X$  in characterizing the direct effect of  $Z$  on  $Y$ . Aside from this feature and the obvious point that Pearl characterizes a notion of “direct effect” rather than a notion of “direct cause,” the main differences between my characterization and his have to do with the background frameworks we assume. At least in Pearl (2000a), he assumes as primitive the notion of a causal mechanism, as represented by an equation or a set of arrows directed from certain variables into another. Pearl then *defines* the notion of an intervention with respect to a graph that correctly specifies the causal mechanisms in the system of interest and then uses the graph and the notion of an intervention to help define the notion of direct effect. By contrast, the notion of intervention that I adopt below (chapter 3, section 1) is not defined with reference to a correct causal graph. Rather than assuming the notion of a causal mechanism as primitive, I attempt to use the notion of an intervention to characterize what it is for there to be a causal mechanism or a direct causal relationship connecting  $X$  to  $Y$ . It is also worth noting that Pearl (2000a, p. 14) does offer a definition of the notion of a Markovian parent, which also may be taken as another characterization of the direct causes of a variable: the Markovian parent of  $X$  is the minimal set of predecessors of  $X$  that render  $X$  independent of all of its other predecessors. Unlike (DC), this characterization assumes the Causal Markov condition, discussed below in section 4.
20. Direct causal relationships are also relative to the range of allowable values taken by variables. Consider the following example from Cooper (1999):  $X_1$  represents *history of smoking* and takes one of three values  $\{\textit{none}, \textit{moderate}, \textit{severe}\}$ ,  $X_2$  represents *chronic bronchitis* and takes values  $\{\textit{absent}, \textit{moderate}, \textit{severe}\}$ , and  $X_3$  represents *cough*, which takes values  $\{\textit{present}, \textit{absent}\}$ . Suppose that relative to these values,  $X_1$  is a direct cause of  $X_2$ , which is in turn a direct cause of  $X_3$ : fixing  $X_2$  at any of its three values, manipulating  $X_1$  does not change  $X_3$ . Suppose now that *chronic bronchitis* is replaced with a coarser-grained variable *chronic bronchitis\** ( $X_2^*$ ), which takes just two values  $\{\textit{absent},$

present}. We now face the question of what it means to fix *chronic bronchitis\** at the value *present*. Assume, as seems reasonable, that this involves fixing *chronic bronchitis\** in such a way that it still is allowed to vary between *moderate* and *severe*. If the value of  $X_2^*$  is fixed in such a way that the finer-grained values of  $X_2$  carry additional information about history of smoking that is not carried by  $X_2^*$  and if this additional information is relevant to  $X_3$  (this would happen if, e.g., severe bronchitis is more likely given severe smoking history and cough more likely given severe history), then  $X_3$  and  $X_1$  will be correlated under manipulation of  $X_1$  with  $X_2^*$  fixed at *present*, and it will appear as though  $X_1$  is a direct cause of  $X_3$ . Thus, whether one variable is a direct cause of another may depend on how fine-grained or coarse-grained we make other variables in the system. Looked at one way, this is unsurprising: we've already acknowledged that direct causal relationships are relative to a variable set and replacing a fine-grained variable with a coarse-grained one amounts to a change in variables.

21. A number of other potential counterexamples to transitivity are discussed in the recent literature, including Hall (2000) and Lewis (2000). Lewis defends transitivity, as does Hall, for what he regards as the "central" notion of causation.
22. The example also illustrates a limitation on representational power of the graph-theoretical framework. From the fact that there is an arrow from  $X$  to  $Z$  and an arrow from  $Z$  to  $Y$ , we cannot infer that  $X$  is either a contributing or total cause of  $Y$ . To make such an inference we would have to know something about the functional relationship connecting  $X$  to  $Z$  and  $Z$  to  $Y$ .
23. A similar analysis is provided in Hitchcock (2001a). I should also add that I do not claim that the example can be dealt with only within a manipulability framework.
24. A possible objection to this claim, raised by Ned Hall in correspondence, is that in ordinary garden-variety cases in which there is a backup cause ( $c_1$  causes  $e$ , but if it had not,  $c_2$  would have caused  $e$ ), the occurrence of  $e$  is not sensitive to whether  $c_1$  occurs, even though  $c_1$  causes  $e$ . Hence (it is argued), the insensitivity of  $E$  to changes in  $B$  cannot be what explains our inclination to judge that  $B$  does not cause  $E$ . A full discussion of this objection must await my treatment of backup causes in section 2.7, but the short response is that the case under discussion differs from cases of backup causation like that just described. In the dog bite example, we have insensitivity along a single causal route and hence (I would argue) a genuine failure of transitivity. By contrast, as we shall see in section 2.7, backup causation involves several distinct causal routes: in the above example, the route by which  $c_1$  affects  $e$  is different from the route by which  $c_2$  would affect  $e$  were  $c_1$  not to occur. My claim above is not that whenever one variable or outcome is insensitive to another, the latter does not cause the former, but rather, the more restricted claim that when there is a single route from one variable  $X$  to another variable  $Y$  and  $Y$  is insensitive to  $X$ , then  $X$  does not cause (i.e., is not causally relevant to)  $Y$ . It seems to me that it is an attraction of the analysis given above that it captures our intuitive sense that the dog bite example is very different from cases involving backup causation and that the insensitivity of the effect to changes in the cause has a very different significance in the two kinds of cases. See Hitchcock (2001a) for a similar argument.
25. See Hitchcock (2001a, 2001b) for further discussion.
26. This condition and the condition (M) immediately below are put forward as a characterization of what it is for  $X$  to be a contributing cause of  $Y$  and not as a

characterization of what it is for  $X$  to cause  $Y$  *along a particular directed path or route*. Judea Pearl (2000a, 2000b) has argued that the italicized notion cannot be captured by the device of freezing off-path variables at some value that is employed in **(M)**. The apparent problem is that when there is an indirect path  $P_1$  between  $X$  and  $Y$  and the only other path  $P_2$  between those variables is direct, there are no variables that are off path  $P_1$  to freeze in order to capture the notion of  $X$ 's being a cause of  $Y$  along  $P_1$ . Pearl argues that to capture the notion of a cause or effect along a causal route, we instead need the notion of *deactivating* a causal route, which corresponds to modifying the operation of a causal factor so that it no longer influences some other variable via a particular path. For example, if it were possible to produce a modified pill in the birth control example that still inhibits pregnancy but has no effect on thrombosis except via its influence on pregnancy, then this would amount to deactivating the direct causal route from the pills to thrombosis.

27. Clark Glymour has drawn my attention to the close connection between this claim and the results in Spirtes et al. ([1993] 2000, chap. 4) concerning rigid indistinguishability. In particular, it can be shown that, assuming the Causal Markov condition (see section 4), any two distinct graphs with the same variable set  $V_i$  can be distinguished by introducing a unique exogenous variable (an intervention variable) for each  $V_i$ . The extended graphs will then imply different conditional independence relations.
28. A minor terminological annoyance is that to assess whether  $X$  causes  $Y$ , other factors that are negatively as well as positively causally relevant to  $Y$  must be controlled for. Thus, if "cause" is restricted to mean "positive cause," it is incorrect to say that the only factors that need to be controlled for are "other causes of  $Y$ ." One needs some additional vocabulary to describe the other factors that need to be controlled for. A more fundamental difficulty is that once one moves away from cause variables that are binary valued, it often becomes unclear what value of the cause variable corresponds to the "absence" of the cause and hence what the state is in comparison with which the "presence" of the cause should raise the probability of the effect.

Another important difference between **TC** and **M**, on the one hand, and **CC**, on the other, is that **CC** imposes what has come to be called a unanimity requirement:  $C$  causes  $E$  if and only if  $C$  raises the probability of  $E$  across *all* background contexts  $K$  (or all situations that are "otherwise causally homogenous with respect to  $E$ ") (Cartwright 1983b, p. 25). Some writers, such as Eells (1991), who follow Cartwright in adopting this unanimity requirement, also make it explicit that the causal notion they are attempting to characterize is the notion of  $C$ 's being a positive causal factor for  $E$  with respect to some population  $P$  of individuals, so that  $C$  must raise the probability of  $E$  for all background circumstances instantiated in  $P$ . As explained above, **TC** and **M** are intended to characterize an individual-level notion of type causation that is not population-relative. But even putting this and other differences aside, what **TC** and **M** require for  $X$  to cause  $Y$  is (in effect) merely that there be *some* (appropriate) background context for which manipulating  $X$  will change the value or the probability distribution of  $Y$ ; it is not required that this be true for *all* background contexts. Thus, in agreement with intuition and contrary to **CC**, **TC** holds that short circuits cause fires even though there is a background context (the absence of oxygen) in which the presence of a short circuit does not raise the probability of fire. Short circuits cause fires because there is also

a background context (presence of oxygen) in which manipulating whether a short circuit occurs changes whether (or the probability of whether) a fire occurs. Similarly, to slightly modify an example from Dupre (1984), smoking will qualify as a cause of lung cancer even if there are possible genotypes for which smoking fails to raise or even lowers the probability of lung cancer. See section 3.7 for additional discussion.

29. I owe to my colleague Alan Hajek the observation that taken literally, **CC** rules out the possibility of the sort of case under discussion in which  $X$  causes  $Z$  which causes  $Y$ . Because  $Z$  causes  $Y$ , it follows from (iii) in **CC** that  $Z$  is in  $\{C_i\}$ . However, it follows from (iv) in **CC** that if  $Z$  is in  $\{C_i\}$ ,  $X$  does not cause  $Z$ . One solution to this difficulty would be to add an additional disjunct to the consequent of (iii) that reads “or  $C$  causes  $D$ .” But this will not help with the more fundamental problems with **CC** discussed above.

In her more recent work, Cartwright (1989b, p. 96) abandons requirement (iv), replacing it with a requirement involving information about singular causal processes. I fully agree with Cartwright that some additional information is needed to distinguish what needs to be controlled for in structures like (2.3.4)–(2.3.5) and (2.3.7)–(2.3.8). However, my suggestion is that what is needed is additional information about direct causal relationships or causal routes at the type-level.

30. This corrects the mistaken claim in Woodward (2001c) that there is no interesting necessary condition formulated in terms of facts about conditional independence relationships for  $X$  to be a contributing cause of  $Y$ .
31. To see that we need to control for (ii) and not just (i), consider the following structure.  $X$  and  $W$  are correlated because of a common cause, but  $X$  does not cause  $W$ .  $W$  is a direct cause of  $Z$ , which in turn is a direct cause of  $Y$ . We wish to determine whether  $X$  is a total cause of  $Y$  with  $Z$  as an intermediate between  $X$  and  $Y$ . If we fail to control for  $W$ , it will appear that  $X$  is a total cause of  $Y$  even when there is no causal link from  $X$  to  $Z$ .
32. A modification of this example may help to drive the point home. Suppose that we add an additional arrow from  $W$  to  $X$  (and additional equation  $X = eW$ ) to figure 2.3.1 and equations (2.3.4)–(2.3.5). Suppose as before that  $a = -bc$ . Then  $X$  is not a total cause of  $Y$ . Instead,  $Z$  is the only total cause of  $Y$ . Although  $W$  is independent of  $Y$  conditional on both  $X$  and  $Z$ ,  $W$  is *not* independent of  $Y$  conditional just on  $Z$ , the only total cause of  $Y$ . Nonetheless,  $W$  is not a total cause of  $Y$ . It would be a mistake to infer from the fact that  $Y$  and  $W$  are dependent conditional on  $Z$  that  $W$  is a total cause of  $Y$ .
33. One possible response, suggested by Ned Hall, is that the very blows that occurred might have occurred with a different momentum; the momentum possessed by the actual blows is not an essential property of them. I find this unconvincing for several reasons. First, it is unclear how to determine which properties are essential to an event and which are not. Second, because by hypothesis any momentum greater than  $m_1$  will produce shattering, the proposed maneuver accomplishes nothing unless one is prepared to argue that any blow with momentum greater than  $m_1$  is the same event as an actual blow with momentum  $m_k$ , no matter how large the difference between  $m_1$  and  $m_k$ . This is very implausible. Third, insofar as the motivation for such an appeal to event essences is that it is a way of responding to a potential counterexample to counterfactual accounts of causation, the appeal to contrastive focus accomplishes the same end and both seems more natural and imports less metaphysical baggage.

34. The notion of invariance is discussed in more detail in chapter 6. I emphasize that, on my view, the claim that there is a causal relationship between *C* and *E* (*C* causes *E*) requires only that there be *some* appropriate relationship between *C* and *E* that is invariant. Exactly which relationship is invariant and under which interventions will vary from case to case.
35. **TC** and **M** would be reductive only if the notion of an intervention can be characterized in noncausal terms, and I argue in chapter 3 that it cannot be.
36. Nor, of course, does reproducibility imply that anyone *knows* how to manipulate the effect variable in such a way as to produce systematically the effect of interest.
37. This may be compared with Lewis's theory of counterfactuals, according to which, in the case described above, the counterfactual "If I had flipped the coin with my right hand, it would have come up heads" is true as long as it is true that I both flipped the coin and it came up heads.
38. Of course, if the probability that instances of *D* will cause instances of *R* is sufficiently low, it may be reasonable to expect that *D* will cause *R* at most once, but this does not mean that *D* can cause *R* only once.
39. See Spirtes et al. ([1993] 2000, chap. 3).
40. I thank Clark Glymour for forcefully insisting on the importance of these issues in correspondence. In effect, what is going on in the case of selection bias is that both *T* and *R* influence the value of a variable *S* that measures presence/absence in the observed experimental sample. If we observe only values for which *S* = *present*, *T* and *R* will be associated, even if they are causally independent. (Two unconditionally independent causes of a common effect will be dependent conditional on the common effect.) The Causal Markov and Faithfulness conditions provide a natural framework for dealing both with cases of this sort and with at least some problems associated with mixtures. For the former, see Spirtes et al. ([1993] 2000, especially pp. 209–51). I will add that although I fully agree with Glymour that, in contrast to the Causal Markov condition, **TC** and **M** give us no practical guidance in dealing with the issues of experimental design just described, I do not think this undercuts the role in clarifying the content of causal claims for which I argue in this chapter.
41. I claim no particular originality for the account that follows, which has been heavily influenced by discussions with Chris Hitchcock and Judea Pearl and by the treatments of actual causation in Hitchcock (2001a), Pearl (2000a), and Halpern and Pearl (2002).
42. Cf. Pearl (2000a, p. 312) for a similar diagrammatic representation of this example.
43. For a more detailed argument in support of a similar conclusion, see Hitchcock (2001a).
44. This is the reaction of Lewis (2000) and Schaffer (2000b), both of whom take this to be a case in which the major's orders causally preempt those of the sergeant. It is my own preanalytic assessment as well. By contrast, Chris Hitchcock suggests that we may instead think of the case as involving causal overdetermination: both the major's and the sergeant's orders cause the corporal to advance, even though it is true that on other occasions, when the major's and sergeant's orders disagree, the major's orders alone cause the corporal's behavior. One consideration in favor of Hitchcock's suggestion is that it is supported by an otherwise plausible account of overdetermination

introduced below. One might, on the basis of this account, conclude that the majority judgment about the trumping example is mistaken. A more radical response, also considered (and tentatively endorsed) below, is that the question of whether the sergeant's orders, in addition to the major's, caused the corporal's advance is a "don't care" question, which need not have a determinate answer.

45. To see that the postulation of more than two values for the variables  $M$  and  $S$  (or at least two values, neither of which is "no orders") is not just an ad hoc device for producing the right answer, but is required to motivate the claim that there is an asymmetry between the major's and the sergeant's orders, suppose that for some reason only two possibilities are available to both the major and the sergeant: each may either order "Advance" or give no orders. Now there are just four possible combinations of causal inputs to  $C$ : for three of them, the corporal advances and for the other he does nothing. Although the example requires more discussion than I can give it here, I believe that in this case, there is no longer any asymmetry between the major and the sergeant, no reason for thinking the case is correctly described as one in which the major's orders trump the sergeant's, and no grounds for denying that *both* the major's and the sergeant's ordering "Advance" are causes of the corporal's advance. The case thus becomes one of symmetric overdetermination, which is treated immediately below.

Jonathan Schaffer disputes this assessment. He claims (personal correspondence) that it should make no difference to our judgment about this example whether the variables  $M$  and  $S$  can take more than two values. On his view, what is relevant is what the laws governing the situation say, and these support the judgment that the major's and not the sergeant's orders are efficacious. I disagree. The most obvious way of reaching this conclusion is to construe the relevant law as having a causal claim built into it; that is, as something like (1) "Whenever the major says 'Advance,' his orders cause the corporal to advance, and even if sergeant's orders are to advance, they don't cause the corporal to advance." However, this construal is question-begging and defeats the purpose of using the law as an independent basis to pick out the major's orders alone as causally efficacious. Alternatively, Schaffer may be thinking of the law as something like (2) "Whenever the major orders 'Advance,' the corporal advances," and supposing that he can read off the causal claim that the major's orders alone are efficacious just from the syntactic form of the law, the inference being something like (*c*) major says advance, (*e*) corporal advances, there is a law (2) linking *c* and *e*, therefore *c* causes *e*. Furthermore, (2) does not link the sergeant's orders to the advance, so the former does not cause the latter. This too is unpersuasive. For one thing, (3) "Whenever the major says 'Advance' and the sergeant says 'Advance,' the corporal advances" has just as good a claim to be regarded as a law. If so, by the same syntactic criterion for when a law makes a causal claim true, we get the result that the major's and the sergeant's orders are causally efficacious. Schaffer (2000b) considers a response like this and rejects it on the grounds that considerations of simplicity support the claim that (2) is a law and (3) is not. I don't think that simplicity can be legitimately used in this way. Assuming that we need a system of laws that cover all the possibilities, (2) must be supplemented by additional generalizations specifying what happens when  $M=0$ ,  $S=1$ , and so on, and similarly for (3). These expanded systems will be



equivalent, differing only syntactically, and hence an appeal to simplicity cannot tell us that the system of laws consists of (2) but not (3). As this example suggests, purely syntactic conceptions of the relationship between laws and causal claims are inadequate; see chapter 4.

46. **AC\*** has the following limitation. Recall the trumping example: both the sergeant and the major order "Advance" ( $S = 1$ ,  $M = 1$ ) and the corporal advances ( $C = 1$ ). In the actual circumstances with  $S = 1$ , the value  $M = 0$  (major gives no orders) is redundant, because under  $M = 0$ , it will still be the case that  $C = 1$ . However, with  $M = 0$ , an intervention that changes the value of  $S$  will change the value of  $C$ , and hence **AC\*** leads to the apparently incorrect conclusion that  $S = 1$  is an actual cause of  $C = 1$ .

Several responses are possible. First, it may be held (cf. n. 44) that the trumping case is really a case of overdetermination rather than preemption; if so, the conclusion that  $S = 1$  is an actual cause of  $C = 1$  is correct after all. Second, one might take the view that **AC** and **AC\*** should, so to speak, be applied lexically: given a situation in which one wishes to determine the actual causes, one first applies **AC**. If **AC** identifies some events as actual causes, one goes no further. It is only if **AC** fails to identify any events as the actual causes of some event of interest (as in cases of completely symmetric overdetermination) that one should turn to **AC\*** to identify causes. Following this procedure we will be led by **AC** to the conclusion that the major's orders are the actual cause of the corporal's advancing. **AC\*** never comes into play, and the conclusion that the sergeant's orders are also a cause is avoided.

47. McDermott (1995) reports the result of an informal poll in which the great majority of respondents judge that in a case of symmetric overdetermination both  $c_1$  and  $c_2$  are causes of  $e$ .
48. It is worth noting that Lewis's treatment of preemption does not agree with the analysis in (**AC**) about what the relevant pattern of counterfactual dependence is. See Hitchcock (2001a, pp. 276ff) for discussion.
49. This point is emphasized by Evelyn Fox Keller (2000) in connection with molecular biology.
50. I don't claim that judgments about serious possibility play exactly the same role in all theories of causation; obviously, such judgments will enter into different theories in different ways. My point is simply that such judgments will need to play *some* role in all plausible theories, so it cannot be a fatal objection to a theory that relies at some points on such judgments. This leaves open the possibility of mounting an argument that the manipulability theory is defective because it relies more heavily on such judgments than other theories or because such judgments enter into the manipulability theory in the wrong way.
51. Because an event that looks disjunctive from the point of view of one choice of representation or choice about which predicates correspond to natural properties will be nondisjunctive from the point of view of another choice.
52. This is consistent with the remarks at the end of 2.3 because our choice of variables will reflect the possibilities we take seriously.

### Chapter 3 Interventions, Agency, and Counterfactuals

1. Suppose that  $T$  is in fact a direct cause of  $R$ , but that some third variable  $Y$  is a direct cause of both  $T$  and  $R$  and that there is cancellation along the direct and indirect paths from  $Y$  to  $R$ , so that  $Y$  is not a total cause of  $R$ . An intervention

$I$  on  $T$  should nonetheless not be correlated with  $Y$ , for if  $Y$  is correlated with  $I$ , the change in  $R$  that results when  $I$  changes  $T$  will confound the influence of  $T$  on  $R$  with the influence of  $Y$  on  $R$ . In requiring that  $I$  not be correlated with contributing causes of  $R$ , such as  $Y$ , we avoid this problem.

2. The strategy of first defining an intervention variable and then using this to define the notion of an intervention is taken from Woodward and Hitchcock (2003). The present characterization attempts to correct some inadequacies in Woodward (1997a, 2000).
3. Although I emphasize that in many cases, causal relationships are underdetermined by independence relationships, I should also note, for the sake of completeness, that there are simple cases in which, given the Causal Markov and Faithfulness conditions, the pattern of independence relations does uniquely determine the causal relationships. See Spirtes et al. ([1993] 2000, chap. 4) for examples and a general discussion of the underdetermination issue.
4. Thus, individual probability claims (e.g., that this coin has probability 0.5 of coming up heads when tossed) do not imply and are not translatable into facts about relative frequencies, but from such probability claims, when combined with other probabilistic assumptions (e.g., that successive tosses of the coin are independent and identically distributed), it follows that certain outcomes, expressible as facts about frequencies, are highly probable.
5. Suppose, for the sake of argument, that there are systems that violate **CM**; for example, suppose that there are structures in which  $C$  is a common cause of  $X$  and  $Y$ , with no causal link from  $X$  to  $Y$  or vice versa, but in which  $X$  and  $Y$  are not independent conditional on  $C$ . In such cases, it would appear that one can still intervene in the sense of **IN** to set  $X$  at various values, and observing no change in  $Y$ , conclude that  $X$  does not cause  $Y$ . The possibility of such cases is one motive for wanting a characterization of intervention that, like **IN**, does not require that the system intervened on satisfy **CM**. On the other hand, once one puts aside cases involving mixing of distinct populations and certain other possibilities, it is far from obvious that there are any real-life cases that violate **CM** (Spirtes et al. [1993] 2000; Hausman and Woodward 1999, forthcoming a). Moreover, as Glymour (personal correspondence) notes, cases like those involving mixed populations can be successfully analyzed using **CM**. Glymour also notes that if, in the above example, one assumes that either  $X$  causes  $Y$ , or if not, that there are other unobserved common causes  $U$  of  $X$  and  $Y$  such that  $X$ ,  $Y$ ,  $C$ , and  $U$  satisfy **CM**, and then does the experiment of manipulating  $X$  and applies the analysis of Spirtes et al., one will reach the correct conclusion about whether  $X$  causes  $Y$ . In other words, assuming **CM** does no harm. Thus, a characterization of interventions that assumes **CM** may be, as a practical matter, completely appropriate for problems of causal inference in experimental contexts. On the other hand, adopting such a characterization seems to foreclose the possibility of using the connection between causation and manipulation to explore the conditions under which it is reasonable to expect **CM** to hold, a project pursued in Hausman and Woodward (1999).
6. This motivation is explicit in Cartwright and Jones (1991).
7. An example: You see a wire running from a switch to a light and wonder whether flipping the switch causes the light to go on and off. You may not know whether this causal claim is true—that is what you want to find out—but it is a very plausible guess that *if* the position of the switch causally affects the light, it does so via the wire. Thus, an experimental manipulation of the switch

that involves severing the wire will not be illuminating for the purposes of determining whether the position of the switch affects the light.

8. Pearl discusses this example on his home page. See Pearl (2001) under “Discussion with readers,” “Can  $do(x)$  represent practical experiments?”
9. I emphasize again that I do not intend these remarks as a criticism of **PI** for the purposes to which Pearl puts it. Because **PI** is characterized by reference to the correct graph for the system under investigation, this will be a graph in which, in the above example, there is an arrow from  $Z$  to  $R$  but no arrow from  $X$  to  $R$ , and in which, as Pearl says, the experimental manipulation manipulates both  $X$  and  $Z$ . Given that this is stipulated to be the graph, the causal effect of  $X$  on  $R$ , as defined by Pearl, will be null, as it should be. My concern is not that Pearl’s account gives the wrong answer in this case, but that **PI** is not a notion of intervention that can be used to characterize what it is for  $X$  to cause  $R$ . Instead, we must presuppose some independent characterization of what it is for  $X$  to cause  $R$  when we use **PI**.
10. I don’t deny that some people say things like “Being a cheetah causes an animal to run fast.” My point is that such claims are unclear and should be replaced with claims that are clearer.
11. This is a common theme in the literature on experimental design. See, for example, Holland (1986) and Rubin (1986).
12. Is there some other way of understanding the claim that causation is “constituted” by our beliefs and attitudes that avoids the conclusion that if these had been different, different causal claims would have been true? It might seem that rigidification is an obvious strategy for avoiding this conclusion: tie causal relationships rigidly to the beliefs and attitudes we presently possess, and the conclusion will no longer follow. However, I agree with Menzies and Price’s judgment, echoing David Lewis (1989), that “this strategy does more to hinder the expression of the worry [about subjectivism] than to make it go away” (1993, p. 199). One way of seeing this is to note that I might equally adopt a parallel strategy of rigidification for concepts defined by reference to what Jim Woodward finds interesting or amusing. The availability of this strategy does not show that these concepts are not “subjective” in some important sense. Similarly for Menzies and Price’s characterization of causation.
13. A very similar view is defended in Hitchcock (1996).
14. Of course, subjectivists will hold that we have different attitudes toward different sorts of correlational claims, depending on whether or not they expect them to continue to hold in the future. But it should be clear from my discussion above that the contrast between those correlations we expect to hold in the future and those we do not does *not* coincide with the contrast between causal and noncausal relationships or even with the contrast between relationships we regard as causal and those we do not.
15. A great deal of the learning that underlies the acquisition of causal concepts involves the acquisition of practical skills and habits that are not in any obvious sense “based on” or derived from conscious experiences. There is now considerable evidence supporting the independence of the systems involved in the acquisition of such “procedural” memories from the “episodic” memories of particular experiences on which classical empiricism is based (see, e.g., Squire and Kandel 1999). For this reason, among others, what is learned should not be equated with what is derived from conscious experience.
16. For additional discussion, see Pearl (2000a, pp. 240ff).

17. For Lewis's quite different treatment of broadly analogous cases involving token causation, see appendix E in 1986d, pp. 193–212.
18. A similar position is defended in Kvart (1986).
19. I am indebted to Alan Hajek for discussions of this example. Hajek believes that a version of the example was discussed in an unpublished lecture by Lewis and perhaps by others as well.
20. Another way of seeing the difference between the interventionist account and Lewis's in connection with examples of this sort is to note that an intervention on  $X$  is always defined with reference to some putative effect  $Y$ . This is necessary if we are to require that the intervention not affect  $X$  independently of  $Y$ , and so on. By contrast, for Lewis, the notion of a closest possible world to be used to assess counterfactuals of the form "If  $X$  had not occurred . . ." makes no reference to  $Y$ .
21. The difficulties posed by this example and the previous one cannot be avoided by altering the criteria (S1–4) in such a way that less importance is assigned to S1, for Lewis would then lose his solution to what he calls "the future similarity objection" to his theory (1986d, p. 43). That is, if S1 is demoted in importance, counterfactuals such as "If Nixon had pushed the button, there would have been no nuclear holocaust" will come out true.
22. For discussion, see Ohanian (1976) and Skyrms (1980).

#### Chapter 4 Causal Explanation: Background and Criticism

1. Recall that I am using "causal explanation" in a broad sense, which encompasses, for example, the sorts of explanations that the *DN* model attempts to capture. Where "explanation" occurs in this and subsequent chapters, what is meant is always causal explanation in this broad sense.
2. For ease of exposition, I ignore the possibility that the underlying explanation has an *IS* structure. This will not affect the points that follow.
3. See, for example, Gardiner (1952) and Popper (1959).
4. I readily acknowledge that the notion of a law being at the same level or using the same vocabulary as a singular causal claim is an unclear and imprecise notion. The argument that follows does not require that this notion be made precise but instead claims only that, however we unpack the notion of a direct generalization, the laws "underlying" most singular causal claims are not direct generalizations of those claims. More generally, the unclarity of the notion of a direct generalization reflects the unclarity of the more general notion of a law's "underlying" a causal claim (see 4.4).
5. In other words, to say, without resorting to the hidden structure strategy, that a generalization such as (4.3.2) can be used to explain even though it is not exceptionless and makes no claim about its consequent having high probability is to opt for a non-*DN/IS* account of explanation. I defend such an account in chapter 5.
6. Many writers claim they relate facts. My own view is that singular causal claims always have a contrastive structure, although this is often implicit. They should be understood as claiming that the occurrence of event  $c$  rather than alternative event  $c'$  caused the occurrence of  $e$  rather than alternative  $e'$  (see Woodward 1984, and chapter 5). Needless to say, if this proposal is correct, it considerably complicates the task of relating singular causal claims to the laws that underlie them.

7. Davidson's criterion is that events are identical if and only if they have exactly the same causes and effects. The proposal is circular or at least unilluminating in the sense that it does not give us a condition that can be used to tell whether two events are identical. If we want to know what the causes and effects of some event *c*, referred to in a singular causal claim, are, we need to know what laws it instantiates, and to know this we need to know which event described in the vocabulary of those laws is identical with *c*. In addition, the proposal makes causes and effects highly "fragile" and their identity dependent on extrinsic or relational facts in an unintuitive way. Jones's death will be a different event depending on whether it is caused by shooter 1 or shooter 2, even if both deaths are in every "intrinsic" respect qualitatively identical and have the same spatiotemporal location. In indeterministic contexts, *c* will be a different event depending on whether or not it causes some probabilistic effect *e*. It is well-known that this sort of fragility creates many difficulties for counterfactual accounts of causation.
8. Davidson requires that laws take this form because of the uniqueness claims that are built into the definite descriptions used in singular causal claims, for example, *the* short circuit caused *the* fire. For an argument that laws of nature rarely take the form indicated in (L), see Woodward (1986).
9. Because a type-causal claim such as "Short circuits cause fires" makes no uniqueness claims, the laws underlying it will need to take a different form from (L).
10. For a more general argument for this contention, see Horgan and Woodward (1985). See also Glymour (1999) for a similar claim.
11. What fundamental laws of physics underlie or ground or are "instantiated" by (4.4.3)? Of course, we can say that if various fundamental physical laws had been sufficiently different, (4.4.3) would not have been true. For example, if the laws of electromagnetism had been sufficiently different, many biochemical reactions and mechanisms essential for carbon-based life to evolve would not occur, oil and human beings would not exist, and there would have been no OPEC. I suspect, however, that it is hard to say anything more precise than this about the relation between (4.4.3) and these underlying laws.
12. Of course, this counterexample can be avoided by building the appropriate causal content into the statement of the law, that is, by taking the law to say that the angle of the sun and the height of the flagpole cause the length of the shadow, but not that the length causes the height. However, we then lose the possibility of using the law to provide an independent account of causal claims.
13. I have in mind semantic theories of the sort developed by Kripke and Putnam. Aside from some brief remarks in Armstrong and Heathcote (1991), who contend that it is an a posteriori necessity that all true singular causal claims instantiate laws in much the same way as it is (allegedly) a necessary but a posteriori truth that water is H<sub>2</sub>O, I know of no attempt to systematically develop this line of thought in the philosophical literature. Presumably, the idea would be that acquaintance with paradigmatic causal interactions (collisions of billiard balls, etc.) fixes the reference of the word "cause," but after empirical investigation we learn that what this word refers to has to do with instantiation of a law, the latter being the essence or hidden structure that underlies causation in something like the same way that molecules of a certain structure turn out to be the hidden structure of water. Although I will not attempt a detailed assessment of this idea it faces many obstacles. "Water" is

a natural kind term; it is plausible that those who use this word use it with the intention of referring to some underlying kind with a distinctive structure that (they recognize) may be different from the phenomenal properties that are typically used to identify “water.” It is far from obvious that “cause” operates like a kind term; indeed, those who run the above argument are *not* taking it to refer to a kind of thing or stuff, but to the instantiation of a complicated relational structure. A related problem arises if the reference fixing is understood in the usual way as involving a causal chain running from the referent, because this commits us to thinking of the instantiation of a law by two events related as cause and effect as itself something that can have causal effects. (For these reasons, I think that a more natural way of developing the Putnam–Kripke strategy would be to take the referent of “cause” to be something more thing-like and less problematically causally efficacious, for example, transfer of energy and momentum. However, as we’ve repeatedly observed, this conception faces its own problems.) Also arguing against the idea that “cause” is a natural kind term with a hidden essence is the fact that, as argued in chapter 2, there are a number of different sorts of causal claims: claims about total causes, contributing causes, and actual causes. Does each correspond to a different hidden structure? Finally, there is what Devitt and Sterelny (1987) call the *qua* problem. In this context, one aspect of the problem can be put this way: suppose that the manipulability theory is right that causal relationships are coextensive with those relationships that potentially support manipulation. Suppose also that all causal relationships instantiate laws. What, then, entitles us to take the referent of “cause” to have to do with instantiation of a law rather than with a relationship that supports manipulation, especially given the unclarity that (I have argued) attach to the former notion?

14. Thus, Scriven is again quite right to insist, against Hempel, that even if it were true that laws had to be cited as evidence for every explanation, it would not follow that the laws must be viewed as part of the explanation.
15. The “intuitions” of many philosophers about this case may be heavily influenced by a prior commitment to the hidden structure strategy and nomothetic models of explanation. Thus, it may help to consider a concrete scenario: your knee hits the desk and the ink spills on the living room rug. Your spouse enters the room and asks, “Why are there ink stains on the carpet?” Will he or she regard it as an explanation if you say, “The laws underlying the occurrence of this event are those of Newtonian mechanics”?
16. Or at least this follows if we accept the principle of the common cause, according to which, if *C* and *E* are correlated, then either *C* causes *E*, *E* causes *C*, or *C* and *E* have one or more common causes.
17. I do not mean to deny that there are features of a claim that are relevant to its explanatory import that may be unknown to users of the claims. To take only the most obvious possibility, the fact that a causal claim is false is certainly relevant to its explanatory import even if this fact is unknown to all who use the claim. However, it is not this sort of possibility that is at issue in the current discussion. We’re assuming that causal claims like (4.2.3) are true and asking whether it is plausible to think of them as explaining in virtue of conveying information that is epistemically hidden from the users of such claims. Readers who think it is not possible for a causal claim to be true unless some law or *DN*-type structure underlies it are reminded (a) of our earlier discussion of the meaning thesis, and (b) that from the fact that such underlying laws or

structure obtain, it doesn't follow that the causal claim explains in virtue of conveying information about them.

18. I argue in chapter 5 that the explanatory import of most singular causal explanations depends on the truth of causal generalizations that are not explicitly represented in such explanations. In the case of an explanation like (4.2.4), the relevant generalizations tell us, for example, that the collision of one massive object (e.g., a knee) with another (e.g., a desk) often will cause the latter to move or vibrate, that less massive objects (e.g., inkwells) placed on a surface can themselves be caused to move off the surface by the movement of the former, that unsupported objects fall vertically downward until they encounter another massive object, that liquids spill when their open containers undergo sudden motion or upending, that when ink comes into contact with absorbing surfaces like paper or carpet, this causes a mark or stain, and so on. (These are among the generalizations to which Scriven himself calls attention in his original description of this example, quoted above.) Unlike the laws of Newtonian mechanics, this sort of general causal knowledge is known to virtually every human being in our culture over the age of four and, as Scriven observes, much of it is equally known to members of prescientific and pre-literate cultures. In invoking such information, we don't face the epistemic puzzle of how information that no one is aware of can nonetheless be contributing to understanding or explanatory import that we encountered in connection with the implicit DN argument interpretation of (4.2.4) considered earlier.

## Chapter 5 A Counterfactual Theory of Causal Explanation

1. It might seem that there is an obvious response to these claims about the inadequacy of (Ex. 5.1.1): that the above remarks conflate two distinct explanations or explanatory tasks. First, there is the task of explaining some singular explanandum of interest, for example, why some particular raven,  $a$ , is black. Second, there is the quite distinct task of explaining a generalization or regularity, in this case, why  $(L_1)$  all ravens are black. The generalization  $(L_1)$  can be used, just as the DN and other nomothetic models claim, to provide a perfectly adequate explanation of why some particular raven is black, although of course we can't use this generalization to explain why all ravens are black. To explain why all ravens are black, we do indeed require something like the explanation involving genetic and biochemical mechanisms described above, but it is a mistake to think that such information is required in order to explain why some particular raven is black. If we feel dissatisfied with (Ex. 5.1.1), this is not because it is no (or a poor) explanation of why the particular raven,  $a$ , is black, but rather because we also want to know why all ravens are black and fail to distinguish sufficiently clearly between these two "levels" of explanation.

There are at least two things that can be said about this response. The first is that even if we accept this "two-level" account, it still seems uncontroversial that an acceptable explanation of why  $(L_1)$  all ravens are black will not be an explanation that (just) exhibits a nomologically sufficient condition for the explanandum  $(L_1)$ , but instead will be an explanation that, by exhibiting the relevant genetic and biochemical mechanisms, answers a range of what-if-things-had-been-different questions regarding the explanandum  $(L_1)$ . Thus,

the *DN* model will not be the correct (or at least will not be a complete) theory of the (upper-level) explanation of regularities like ( $L_1$ ), even if it is the right theory about the (lower-level) explanation of singular explananda via subsumption under regularities. The “two-level” approach thus has the implication that there are two quite different varieties of explanation that work according to different principles, one having to do with explanation of regularities and the other having to do with the explanation of singular explananda. As suggested in chapter 1, we should adopt such a position only if more monistic approaches have been tried and found wanting. A more natural position is that both the explanation of regularities and the explanation of singular explananda work according to fundamentally the same principles. The theory I present embodies this position.

Quite apart from this, the two-level conception does not reflect the realities of scientific practice. A number of philosophers have argued that the explanation of particular events (or of explananda that are singular sentences) plays little if any role in science. Although I think this is an overstatement, it does seem that in many areas of science, explanations typically or commonly take the form of explanations of generalizations or regularities, with the explanation of particular outcomes being parasitic on this activity, in the sense that it draws on the same information. Thus, one finds in scientific treatises and textbooks explanations of why simple pendulums have periods  $T = 2\pi\sqrt{\ell/g}$ , of why monopolies are output restrictors, of Boyle-Charles law, of Bernoulli’s equation, and of the expression relating pressure and volume in gasses undergoing an adiabatic process. One does not find in addition to these explanations a distinct kind of explanation in which, for example, the period of some particular simple pendulum is explained in terms of the generalization “All simple pendulums of such and such length have such and such period,” or in which the output-restricting behavior of some monopolistic firm is explained by reference to the generalization “All monopolistic firms are output restrictors.” Instead, the explanations of these singular explananda are the same as the explanations of the generalizations of those explananda. Similarly, the explanation of why some particular raven is black will appeal to the same generalizations, mechanisms, and so on that might be used to explain why all ravens are black.

2. For an earlier and, in some respects, more detailed statement of the ideas defended in this section, see Woodward (1984).
3. These ideas are developed in more detail in Woodward (1993a).
4. For additional discussion of this “eliminative” strategy in connection with singular-causal explanation, see Woodward (1990).
5. For reasons described in Kvart (2001).
6. I thank Richard Healey for a very helpful conversation about examples of this sort.
7. On my view, the firing is a cause of the bombing, but I agree with Hall that this example lacks a feature that is present in many paradigmatic examples of “productive” causation. Although I lack the space for detailed discussion here, I believe that this feature does not have to do with spatiotemporal continuity, but rather with the relative instability or lack of invariance of the counterfactual relationship between Billy’s firing and the bombing: although the latter event counterfactually depends on the former in the actual circumstances, there are many small perturbations in those circumstances (e.g., changes in



the enemy's intentions, presence of other threats to Suzy, etc.) under which this relationship would not continue to hold. In other words, my suggestion would be that "production" can be understood as relatively invariant causation, where "causation" is understood in terms of interventionist counterfactuals. See chapter 6, section 14 for brief additional discussion and Woodward (forthcoming a) for a more detailed treatment.

8. For more detailed discussion, see Woodward (2002b).
9. Suggested by Alan Hajek.
10. For a similar line of argument, see Garfinkel (1981).
11. There is an additional point illustrated by the examples that is worth making explicit. Some laws or generalizations have explanatory or causal asymmetries built into them: the law by itself fixes the direction of explanation. Thus, in the law  $F = ma$ , explanation always flows from the force incident on an object to its acceleration and not vice versa. However, as the gas law examples show, not all laws are like this: some laws do not come with a prespecified direction of explanation. Given the right conditions, one can explain the value of any one of the variables  $V$ ,  $P$ , and  $T$  in the ideal gas law in terms of the other two. As the above examples illustrate, the directional or asymmetric features of explanations involving the ideal gas law derive instead from the details of the particular system or experimental setup to which the law is applied. If the system is such that volume is fixed, then we cannot use the ideal gas law to explain the volume, although, depending on the details of the experimental arrangement, we may be able to appeal to the volume to help explain the values of  $P$  or  $T$ . By contrast, with a different experimental setup and a variable volume, explanation may well run in the opposite direction. A similar point is made by Dan Hausman (1998).

## Chapter 6 Invariance

1. Thus, I count generalizations such as the ideal gas law and the gravitational inverse square law as invariant generalizations over some interventions and changes in background conditions, even though those generalizations are not "exactly" true in any circumstances.
2. For discussion, see, for example, Fodor (1991), Hausman (1992), Kincaid (1989), and Pietroski and Rey (1995).
3. At the risk of belaboring the obvious, a note on terminology is perhaps appropriate here. Obviously, the initial or boundary conditions that play a role in some explanations will be described by generalizations, some of which will be noncausal or accidental. The point that I am making above is not that such accidental generalizations play no role in explanations. What I mean by a generalization having "explanatory import" or "figuring in an explanation" is that the generalization itself describes a causal relationship that connects initial conditions and the explanandum in a way that satisfies the requirements in chapter 5. I assume that every explanation must appeal to at least one generalization that describes an explanatory relationship. The notion of invariance is meant to characterize the common feature that such generalizations will possess.
4. A classic source for the role of symmetry and invariance conditions in the characterization of physical laws is Wigner (1967). Among philosophers, this connection is emphasized by Skyrms (1980) and van Fraassen (1989).

I caution, however, that Skyrms and van Fraassen give the connection a subjectivist interpretation that I do not endorse; see section 6.12.

5. The question of whether stability under changes in background conditions is necessary for a generalization to qualify as explanatory is addressed later in this section.
6. Indeed, Healey himself goes on to formulate such a connection, which he calls “internal robustness.” It differs from the connection that I defend in several respects, including its all-or-nothing character.
7. These remarks bear on a notorious difficulty for the *DN* model of explanation. The problem, as noted by Hempel and Oppenheim ([1948] 1965, p. 273) in their famous footnote 33, is how to distinguish genuine explanations of generalizations, such as a derivation of (an approximation of) Galileo’s law of free fall from Newton’s laws, from spurious explanations, such as a derivation of Galileo’s law from the conjunction of Galileo’s law and the Boyle–Charles law. The account developed above provides a basis for this distinction: Newton’s laws are invariant under testing interventions that would alter Galileo’s law, whereas the conjunctive Galileo–Boyle–Charles law is not. That is, interventions on the values of variables figuring in Newton’s laws, such as the mass and radius of the earth, would result in various alternatives to Galileo’s law. Newton’s laws show how the truth of Galileo’s law depends on the values of certain variables: we are shown how the generalization governing the behavior of falling bodies would have been different if the values of these variables had been different. By contrast, there is no intervention on the value of a variable figuring in the conjunctive Galileo–Boyle–Charles law that would lead to Galileo’s law being false. The conjunctive law in no way shows what the truth of Galileo’s law depends on or under what conditions alternatives to Galileo’s laws would result. Newton’s laws thus explain and the conjunctive law fails to explain why Galileo’s law holds, according to the conception of explanation I have been defending. See Hitchcock and Woodward (2003) for additional discussion.
8. See Phillips (1958), and for relevant discussion, Hoover (1988).
9. For additional discussion of this example as well as some other considerations relevant to explanatory depth, see Hitchcock and Woodward (2003).
10. What does “large” mean? I will not try to make this notion precise. Indeed, for reasons that will emerge below, I am skeptical that it can be made precise. Clarifying this notion is important for those who think that scope is important for explanatory status, but the view I defend below is that scope is irrelevant to explanatory status.
11. Systems that represent exceptions to a generalization must, of course, be distinguished from systems to which the generalization merely fails to apply. I follow Pietroski and Rey (1995) in thinking of exceptions as involving cases in which the behavior of a system satisfies the antecedent of a generalization but not its consequent. By contrast, a generalization will fail to apply to a system if it fails to satisfy the conditions specified in its antecedent. I also remind the reader that approximately correct generalizations can count as invariant; thus, when I speak of “exceptions,” I mean cases in which the antecedent of a generalization is satisfied or approximately so, and its consequent is not even approximately satisfied.
12. For example, this is the *de facto* position adopted in Fodor (1991) and Pietroski and Rey (1995).

13. One possible response to the examples that follow is to try to develop a more precise notion of support, according to which the counterfactuals described below are not really supported by the generalizations associated with them. I shall not explore this alternative. My suspicion is that, if developed adequately, it will coincide with the position I defend.
14. Recall the distinction between conditioning and intervening discussed in chapter 2. (6.9.4) and (6.9.6) seem true if interpreted as claims about what it is reasonable to believe, given certain background information, but not as claims about what would happen if a Sisley were introduced into room 17 or if penny were introduced into Clinton's pocket as a result of an intervention.
15. For criticism of the idea that there is a domain-independent notion of simplicity that we can use to choose among competing theories, see Sober (1988, chap. 2).
16. In addition to Mitchell, see especially Cooper (1998), who explicitly associates resiliency in Skyrms's sense with my notion of invariance. Skyrms has argued for a similar view in correspondence and conversation.
17. Similarly, to take an example Skyrms draws from Ayer (1956), the fundamental reason we do not think the generalization "All the cigarettes in this case are made of Virginia tobacco" is a law of nature has to do with its noninvariance rather than its nonresilience: as both Ayer and Skyrms note, this generalization is not invariant under an intervention that consists in filling the case with Turkish cigarettes.
18. For example, aside from Beatty, the other participants in the PSA 96 symposium "Are There Laws of Biology?" (Brandon 1997; Sober 1997; Mitchell 1997) all tie issues about the existence of laws in biology to issues about biological explanation.
19. For additional discussion of *ceteris paribus* laws, see Woodward (2002a).
20. Exactly what, if anything, one must know about the completer is typically left unclear in completer accounts.
21. A natural question is whether there are plausible additional constraints that, when added to the version of the completer account described above, will avoid these difficulties. I cannot canvas all the possibilities here, but some brief remarks on the additional constraints suggested in the literature will help to support my claim that the difficulties are not easily avoided. Pietroski and Rey (1995, p. 90) add the constraint that the completer must be "independent," where this is understood in such a way as to "exclude factors whose only explanatory role is to save a proposed [*ceteris paribus*] law." They also require that the completer must explain other things besides the failure of the law. Their discussion conflates two distinct issues: whether the completer is independent in the sense just specified, and the epistemic issue of whether one knows how to independently describe the completer in a way that is not tantamount to saying that the law holds except when it doesn't. The examples described above do involve completers that are independent in the nonepistemic sense described in the quoted passage and hence are counterexamples to their proposal. Hausman (1992, p. 137) claims that, "when one takes an inexact generalization to be an explanatory law, one supposes that the *ceteris paribus* clause picks out some predicate that when added to the antecedent of the unqualified generalization makes it an exact law." This formulation appears to be straightforwardly subject to the counterexamples given above. Hausman (pp. 140ff) also describes an additional set of conditions that must be

satisfied for it to be reasonable to believe that an inexact *ceteris paribus* generalization is completable into an exact law. The conditions appear to be unnecessary if, as I have argued, the thesis of macrodeterminism by itself gives one very general reasons to believe that there must be a completer. Moreover, the conditions themselves seem problematic. One is that the “modifications or qualifications of the theory that make [a candidate *ceteris paribus* generalization] more reliable not be ad hoc” (p. 141). But whether the completer for a candidate *ceteris paribus* law is simple or complex, ad hoc or non-ad hoc, appears to have little to do with whether it is invariant at all, and if so, over what range of interventions, and hence little to do with whether it can be used to explain. Many generalizations in the special sciences probably have very complex and ad hoc completers, but this fact by itself does not make them unexplanatory.

## Chapter 7 Causal Interpretation in Structural Models

1. A more detailed discussion of many of the ideas in this chapter, including additional examples and applications, as well as a comparison with alternative concepts of causation such as Granger causation can be found in Woodward (1995, 1997a, 1999).
2. Suppose that we are attempting to estimate the value of a parameter  $\vartheta$  and we have a sample of  $n$  observations  $y_1 \dots y_n$ . An estimator  $t$  for  $\vartheta$  will be a function  $t(y_1 \dots y_n)$ —itself a random variable—of these observations and will be unbiased for  $\vartheta$  if its expected value,  $E(t) = \vartheta$ . Linear estimators are those for which  $t$  is a linear function, that is,  $t = c_1y_1 + c_2y_2 + \dots + c_ny_n$ , where  $c_1 \dots c_n$  are constants. The estimator among the class of unbiased, linear estimators that has a minimum variance will be the best linear unbiased estimator (BLUE).
3. I thank Clark Glymour for helpful correspondence regarding this point.
4. For an attempt to understand the error term purely correlationally, see Papineau (1991) and for more detailed discussion, see Woodward (1999).
5. Freedman regards the distributional invariance conditions as “more plausible” than the value invariance assumptions (forthcoming, p. 3). My contrary view is that if we are prepared to believe that the mechanism that generates equation (7.1.7) is fully deterministic—that is, that (7.1.7) is to be taken literally as a description of the behavior of the individuals in the population to which it applies and that the mechanism generating the values of the error term is similarly deterministic—then the value invariance assumptions look very plausible and are perhaps even required. Suppose, for example, that the error term happens to take the value  $u_i$  when the variable  $X_i$  is set by an intervention to the value  $x_i$  on some particular occasion. If the mechanism governing the error term is deterministic and if, as we have been assuming, there is no causal connection running from  $X_i$  to  $U$ , it seems very natural to suppose that if  $X_i$  had instead been set on that occasion to some other value  $x_i^*$ , the value of  $U$  would still have been  $u_i$ . My inclination is to think that if we are not willing to accept this counterfactual, that is because we do not really believe that the mechanism governing the error term acts deterministically. If this mechanism is indeterministic, then I agree that there is no fact of the matter about what the value of the error term would have been had  $X_i$  been set to a different (or, for that matter, the same) value.

6. Philosophers who have claimed that the satisfaction of the uncorrelatedness assumption is necessary for a regression equation to have a causal interpretation or that this assumption is in some way central for causal interpretability include Irzik (1996) and Papineau (1991). The statisticians, econometricians, and contributors to the causal modeling literature who share this assumption are legion; Cooley and LeRoy (1985) provide a recent statement. See Woodward (1999) for more detailed discussion with examples.
7. At the risk of belaboring the obvious, let me be explicit about what is being envisioned. The notion of an intervention was defined in chapter 3 with respect to the true causal structure in which the variable intervened on is embedded. In this sense, intervention is not a representation relative notion. By contrast, modularity is a feature of our representations, of a system of equations, directed graph, or some other device for representing causal relationships. When I speak of an intervention on  $Y$  disrupting or changing some other equation in which  $Y$  does not figure (e.g., (7.4.2)), I mean that from the perspective of some particular system of equations (e.g., (7.4.1)–(7.4.2)), the upshot of the intervention is that the functional relationship described by that equation, in this case, the relationship between  $X$  and  $Z$ , changes. Of course, presumably, what is really going on is that the system (7.4.1)–(7.4.2) is misspecified in some way. For example, there may be a causal relationship from  $Y$  to  $Z$  that is left out of (7.4.1)–(7.4.2). The intervention on  $Y$  doesn't change the real causal relationship between  $X$  and  $Z$ : it just looks that way from the perspective of (7.4.1)–(7.4.2).
8. In a series of papers, Nancy Cartwright (2001, 2002) has objected to the claim that a system of equations that adequately and completely represents a set of causal relationships should be modular. Cartwright claims that the “job” that a system of equations is designed to do, which she takes to be (a) that of representing causal relationships, should be distinguished from the quite different job of (b) representing relations that can be interfered with separately, as Modularity requires. According to Cartwright, we “change the subject” away from (a) when we impose requirement (b). She writes that “the equations are not supposed to give information . . . about what equations might hold if they did not hold,” and it is clear from her subsequent discussion that she means this objection to apply in particular to the idea that a system of equations should be understood as conveying information about what would happen to the other equations under circumstances in which one of the equations is disrupted by an intervention.  
 The very different point of view assumed in the manipulationist account is that what is required for a set of equations fully and accurately representing causal relationships (job (a)) is that it also do (b); that is, it is assumed that (a) and (b) are not distinct jobs at all. For additional discussion, see Hausman and Woodward (forthcoming a and forthcoming b).
9. My discussion in this section has focused entirely on *recursive* systems of equations: systems in which there are no causal loops or cycles in which  $X$  causally influences  $Y$  and  $Y$  in turn causally influences  $X$ . The notion of modularity and the other ideas discussed in this section readily generalize to nonrecursive systems; for discussion, see Woodward (1999).
10. Additional rules governing the set operator and a much more detailed and systematic discussion connecting it to graph-theoretical representations of causal relationships are given in Pearl (2000a, pp. 85ff.). Like **PM**, **PLI**, **PM2**,

and **PM3**, such rules connect the set operator (and the notions of intervention and invariance) to the probability calculus and to more familiar mathematical operations such as conditioning, and they also make explicit when we may move from causal knowledge and information derived from passive observations (recorded in the probability distribution  $Pr$ ) to predictions about what will happen under interventions.

11. It seems at least logically possible for this assumption to be violated. Imagine a case in which intervening to change the value of  $C$  always changes the value of  $E$  and in which there is no other cause of  $E$  besides  $C$  but in which the precise quantitative value of the conditional probabilities  $Pr(E/C)$ ,  $Pr(E/-C)$  varies from occasion to occasion. Causes of this sort are called “irregular” in Woodward (1993a). Many cases of irregularity arise because the specification of a cause “averages” over heterogeneous microstates. It appears to be logically possible that there are causes that are, so to speak, irreducibly irregular, but I know of no uncontroversial examples.

## Chapter 8 The Causal Mechanical and Unificationist Models of Explanation

1. One might well wonder what the basis for this judgment is. Can't the gas as a whole be “marked” (e.g., by heating it), and won't the gas transmit this mark, at least for awhile?
2. For a convincing and much more detailed discussion of the differences among the various activities that fall under the rubric of unification and for historical examples illustrating the point that formal or mathematical unification often has little to do with explanation, see Morrison (2000).
3. Salmon (1989) makes a similar observation in his discussion of Michael Friedman's account of explanatory unification.

## Afterword

1. I'm grateful to Glymour for insisting in correspondence on the importance of this element in unification and regret that I do not have more to say about it.

## References

- Achen, C. 1982. *Interpreting and Using Regression*. Beverly Hills: Sage.
- Alderich, J. 1989. "Autonomy." *Oxford Economic Papers* 41: 15–34.
- Anscombe, G. [1971] 1993. "Causality and Determination." Reprint, *Causation*, ed. E. Sosa and M. Tooley. Oxford: Oxford University Press.
- Armstrong, D., and A. Heathcote. 1991. "Causes and Laws." *Nous* 25: 63–74.
- Ayer, A. 1956. "What Is a Law of Nature?" *Revue Internationale de Philosophie* 10: 144–65.
- Baillergon, R., L. Kotovsky, and A. Needham. 1995. "The Acquisition of Physical Knowledge in Infancy." In *Causal Cognition: A Multidisciplinary Debate*, ed. D. Sperber, D. Premack, and A. Premack. Oxford: Clarendon Press.
- Barkan, D. 1999. *Walter Nernst and the Transition to Modern Physical Science*. Cambridge, UK: Cambridge University Press.
- Barnes, E. 1992. "Explanatory Unification and the Problem of Asymmetry." *Philosophy of Science* 59: 558–71.
- Beatty, J. 1979. *Traditional and Semantic Accounts of Evolutionary Theory*. Ph.D. diss., University of Michigan, Ann Arbor. University Microfilms International.
- Beatty, J. 1995. "The Evolutionary Contingency Thesis." In *Concepts, Theories and Rationality in the Biological Sciences*, ed. J. Lennox and G. Wolters, 45–81. Pittsburgh: University of Pittsburgh Press.
- Bennett, J. 1984. "Counterfactuals and Temporal Direction." *Philosophical Review* 93: 57–91.
- Bennett, J. 1987. "Event Causation: The Counterfactual Analysis." In *Philosophical Perspectives 1: Metaphysics*, ed. J. Tomberlin. Atascadero, CA: Ridgeview.
- Bogen, J., and J. Woodward. 1988. "Saving the Phenomena." *Philosophical Review* 97: 303–352.
- Brandon, R. 1997. "Does Biology Have Laws?" In *PSA96, S444–S457*, ed. L. Darden. East Lansing, MI: Philosophy of Science Association.
- Bromberger, S. 1966. "Why Questions." In *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, ed. R. Colodny. Pittsburgh: University of Pittsburgh Press.
- Carroll, J. 1994. *Laws of Nature*. Cambridge, UK: Cambridge University Press.
- Cartwright, N. [1979] 1983a. "Causal Laws and Effective Strategies." Reprint, *Nous* 13: 419–37.
- Cartwright, N. 1983b. *How the Laws of Physics Lie*. Oxford: Clarendon Press.

- Cartwright, N. 1989a. "A Case Study in Realism: Why Econometrics Is Committed to Capacities." In *PSA 1988*, vol. 2, ed. A. Fine and J. Leplin, 190–97. East Lansing, MI: Philosophy of Science Association.
- Cartwright, N. 1989b. *Nature's Capacities and Their Measurement*. Oxford: Clarendon Press.
- Cartwright, N. 1995. "Probabilities and Experiments." *Journal of Econometrics* 67: 47–59.
- Cartwright, N. 2001. "Modularity: It Can—and Generally Does—Fail." In *Stochastic Causality*, ed. M. Galavotti, P. Suppes, and D. Costantini, 65–84. Stanford: CSLI.
- Cartwright, N. 2002. "Against Modularity, the Causal Markov Condition and Any Link between the Two: Comments on Hausman and Woodward." *British Journal for the Philosophy of Science* 53: 411–453.
- Cartwright, N. 2003. "Two Theorems on Invariance and Causality." *Philosophy of Science* 70: 203–224.
- Cartwright, N., and M. Jones. 1991. "How to Hunt Quantum Causes." *Erkenntnis* 35: 205–31.
- Churchland, P. 1989. "On the Nature of Explanation: A PDP Approach." In *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, ed. P. Churchland, 197–230. Cambridge, MA: MIT Press.
- Coleman, J., and H. Hoffer. 1987. *Public and Private High Schools*. New York: Basic Books.
- Collingwood, R. 1940. *An Essay on Metaphysics*. Oxford: Clarendon.
- Collins, J. 2000. "Preemptive Preemption." *Journal of Philosophy* 97: 223–34.
- Cook, T., and D. Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.
- Cooley, T., and S. LeRoy. 1985. "Atheoretical Macroeconomics: A Critique." *Journal of Monetary Economics* 16: 283–308.
- Cooper, G. 1998. "Generalizations in Ecology: A Philosophical Taxonomy." *Biology and Philosophy* 13: 555–86.
- Cooper, G. 1999. "An Overview of Representation and the Discovery of Causal Relationships Using Bayesian Networks." In *Computation, Causation and Discovery*, ed. C. Glymour and G. Cooper. Cambridge, MA: MIT Press.
- Cornfield, J., W. Haenszel, and E. Hammond. 1959. "Smoking and Lung Cancer: Recent Evidence and Discussion of Some Questions." *Journal of the National Cancer Institute* 22: 173–203.
- Cowie, F. 1999. *What's Within? Nativism Reconsidered*. New York: Oxford University Press.
- Davidson, D. 1967. "Causal Relations." *Journal of Philosophy* 64: 691–703.
- Devitt, M., and K. Sterelny. 1987. *Language and Reality: An Introduction to the Philosophy of Language*. Cambridge, MA: MIT Press.
- Dickenson, A., and D. Shanks. 1995. "Instrumental Action and Causal Representation." In *Causal Cognition*, ed. D. Sperber, D. Premack, and A. Premack. Oxford: Oxford University Press.
- Dowe, P. 2000. *Physical Causation*. Cambridge, UK: Cambridge University Press.
- Dretske, F. 1977. "Referring to Events." *Midwest Studies in Philosophy* 2: 90–99.
- Ducasse, C. 1926. "On the Nature and Observability of the Causal Relationship." *Journal of Philosophy* 23: 57–68.
- Dummett, M. 1964. "Bringing About the Past." *Philosophical Review* 73: 338–59.
- Dunn, L. C. 1957. "Evidence of Evolutionary Forces Leading to the Spread of Lethal Genes in Wild Populations of House Mice." *Genetics* 43: 157–63.



- Dupre, J. 1984. "Probabilistic Causality Emancipated." *Midwest Studies in Philosophy* 9: 169–75.
- Earman, J. 1993. "In Defense of Laws: Reflections on Bas van Fraassen's *Laws and Symmetry*." *Philosophy and Phenomenological Research* 53: 413–19.
- Eells, E. 1991. *Probabilistic Causality*. Cambridge, UK: Cambridge University Press.
- Eells, E., and E. Sober. 1983. "Probabilistic Causality and the Question of Transitivity." *Philosophy of Science* 50: 35–57.
- Elman, J., E. Bates, M. Johnson, A. Karmiloff-Smith, D. Parisi, and K. Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.
- Fiorina, M. 1995. "Rational Choice, Empirical Contributions and the Scientific Enterprise." In *The Rational Choice Controversy: Economic Models of Politics Reconsidered*, ed. J. Friedman. New Haven: Yale University Press.
- Fodor, J. 1991. "You Can Fool Some of the People All of the Time, Everything Else Being Equal: Hedged Laws and Psychological Explanation." *Mind* 100: 19–34.
- Frautschi, S., R. Olenick, T. Apostol, and D. Goodstein. 1986. *The Mechanical Universe: Mechanics and Heat, Advanced Edition*. Cambridge: Cambridge University Press.
- Freedman, D. 1997. "From Association to Causation via Regression." In *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, ed. V. McKim and S. Turner, 113–61. Notre Dame, IN: University of Notre Dame Press.
- Freedman, D. Forthcoming. "On Specifying Graphical Models for Causation and the Identification Problem." Technical Report 601. Department of Statistics, University of California, Berkeley.
- Friedman, J., ed. 1995. *The Rational Choice Controversy: Economic Models of Politics Reconsidered*. New Haven: Yale University Press.
- Friedman, M. 1974. "Explanation and Scientific Understanding." *Journal of Philosophy* 71: 5–19.
- Frisch, R. [1938] 1995. "Autonomy of Economic Relations." Reprint, *The Foundations of Econometric Analysis*, ed. D. F. Hendry and M. S. Morgan. Cambridge, UK: Cambridge University Press.
- Garcia, J., F. Ervin, and R. Koelling. 1966. "Learning with Prolonged Delay of Reinforcement." *Psychonomic Science* 5: 121–22.
- Gardiner, P., ed. 1952. *Theories of History*. New York: Free Press.
- Gardiner, P. 1959. *The Nature of Historical Explanation*. Oxford: Oxford University Press.
- Garfinkel, A. 1981. *Forms of Explanation: Rethinking the Questions in Social Theory*. New Haven: Yale University Press.
- Gasking, D. 1955. "Causation and Recipes." *Mind* 64: 479–87.
- Gibbard, A., and W. Harper. 1976. "Counterfactuals and Two Kinds of Expected Utility." In *Ifs*, ed. W. Harper, R. Stalnaker, and G. Pearce. Dordrecht: Reidel.
- Giere, R. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.
- Giere, R. 1999. *Science without Laws*. Chicago: University of Chicago Press.
- Glymour, C. 1980. "Explanations, Tests, Unity, and Necessity." *Nous* 14: 31–50.
- Glymour, C. 1999. "A Mind Is a Terrible Thing to Waste—Critical Notice: Jagwon Kim, *Mind in a Physical World*." *Philosophy of Science* 66: 455–71.
- Gopnik, A., C. Glymour, D. Sobel, L. Schulz, T. Kushnir, and D. Danks. Forthcoming. "A Theory of Causal Learning in Children: Causal Maps and Bayes' Nets." *Psychological Review*.

- Green, D., and I. Shapiro. 1995. "Reflections on Our Critics." In *The Rational Choice Controversy: Economic Models of Politics Reconsidered*, ed. J. Friedman. New Haven: Yale University Press.
- Griffiths, A., J. Miller, D. Suzuki, R. Lewontin, and W. Gelbart. 1996. *An Introduction to Genetic Analysis*. New York: W. H. Freeman.
- Haavelmo, T. 1944. "The Probability Approach in Econometrics." *Econometrica* 12 (Supplement): 1–118.
- Hall, N. 2000. "Causation and the Price of Transitivity." *Journal of Philosophy* 97: 198–222.
- Hall, N. Forthcoming. "Two Concepts of Causation." In *Causation and Counterfactuals*, ed. J. Collins, N. Hall, and L. Paul. Cambridge, MA: MIT Press.
- Halpern, J., and J. Pearl. 2000. *Causes and Explanations: A Structural Model Approach*. Technical report R-266, Cognitive Systems Laboratory. Los Angeles: University of California.
- Hausman, D. 1992. *The Inexact and Separate Science of Economics*. Cambridge, UK: Cambridge University Press.
- Hausman, D. 1998. *Causal Asymmetries*. Cambridge, UK: Cambridge University Press.
- Hausman, D., and J. Woodward. 1999. "Independence, Invariance, and the Causal Markov Condition." *British Journal for the Philosophy of Science* 50: 521–83.
- Hausman, D., and J. Woodward. Forthcoming a. "Modularity and the Causal Markov Condition: A Restatement." *British Journal for the Philosophy of Science*.
- Hausman, D., and J. Woodward. Forthcoming b. "Manipulation and the Causal Markov Condition."
- Healey, R. 1992. "Discussion: Causation, Robustness, and EPR." *Philosophy of Science* 59: 282–92.
- Healey, R. 1994. "Nonseparable Processes and Causal Explanation." *Studies in History and Philosophy of Science* 25.3 (June): 337–74.
- Hempel, C. 1965a. "Aspects of Scientific Explanation." In *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, 331–496. New York: Free Press.
- Hempel, C. 1965b. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.
- Hempel, C., and P. Oppenheim. [1948] 1965. "Studies in the Logic of Explanation." Reprint, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, by C. Hempel, 245–90. New York: Free Press.
- Hesslow, G. 1976. "Two Notes on the Probabilistic Approach to Causality." *Philosophy of Science* 43: 290–92.
- Hitchcock, C. 1993. "A Generalized Probabilistic Theory of Causal Relevance." *Synthese* 97: 335–64.
- Hitchcock, C. 1995. "Discussion: Salmon on Explanatory Relevance." *Philosophy of Science* 62: 304–20.
- Hitchcock, C. 1996. "Causal Decision Theory and Decision-Theoretic Causation." *Nous* 30: 508–26.
- Hitchcock, C. 2001a. "The Intransitivity of Causation Revealed in Equations and Graphs." *Journal of Philosophy* 98: 273–99.
- Hitchcock, C. 2001b. "A Tale of Two Effects." *Philosophical Review* 110: 361–96.
- Hitchcock, C., and J. Woodward. 2003. "Explanatory Generalizations, Part 2: Plumbing Explanatory Depth." *Nous* 37: 181–99.
- Holland, P. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81: 945–60.

- Holland, P. 1995. "Comments on David Freedman's Paper." *Foundations of Science* 1: 49–57.
- Hoover, K. 1988. *The New Classical Macroeconomics*. Oxford: Basil Blackwell.
- Hoover, K. 2001. *Causality in Macroeconomics*. Cambridge, UK: Cambridge University Press.
- Horgan, T., and J. Woodward. 1985. "Folk Psychology Is Here to Stay." *Philosophical Review* 94: 197–226.
- Horwich, P. 1987. *Asymmetries in Time: Problems in the Philosophy of Science*. Cambridge, MA: MIT Press.
- Hughes, R. I. G. 1993. "Theoretical Explanation." In *Midwest Studies in Philosophy*, Vol. 18: *Philosophy of Science*, ed. P. French, T. Uehling, and H. Wettstein. Notre Dame, IN: University of Notre Dame Press.
- Humphreys, P. 1989. *The Chances of Explanation*. Princeton: Princeton University Press.
- Irzik, G. 1996. "Can Causes Be Reduced to Correlations?" *British Journal for the Philosophy of Science* 47: 259–70.
- Johnston, J. 1992. "Econometrics: Retrospect and Prospect." In *The Future of Economics*, ed. J. Hey. Oxford: Blackwell.
- Keller, E. 2000. "Making Sense of Life: Explanation in Developmental Biology." In *Biology and Epistemology*, ed. R. Creath and J. Maienschein, 244–60. Cambridge, UK: Cambridge University Press.
- Kim, J. 1999. "Hempel, Explanation, Metaphysics." *Philosophical Studies* 94: 1–20.
- Kincaid, H. 1989. "Confirmation, Complexity and Social Laws." In *PSA 1988*, ed. A. Fine and J. Leplin. East Lansing, MI: Philosophy of Science Association.
- Kitcher, P. 1989. "Explanatory Unification and the Causal Structure of the World." In *Scientific Explanation*, ed. P. Kitcher and W. Salmon, 410–505. Minneapolis: University of Minnesota Press.
- Kvart, I. 1986. *A Theory of Counterfactuals*. Indianapolis: Hackett.
- Kvart, I. 2001. "Lewis' Causation as Influence." *Australasian Journal of Philosophy* 79: 409–21.
- Kyburg, H. 1965. "Comment." *Philosophy of Science* 32: 147–51.
- Leslie, J., and S. Keeble. 1987. "Do Six-Month-Old Infants Perceive Causality?" *Cognition* 25: 265–88.
- Lewis, D. 1973. *Counterfactuals*. Cambridge, MA: Harvard University Press.
- Lewis, D. 1986a. "Causal Explanation." In *Philosophical Papers*, vol. 2. Oxford: Oxford University Press.
- Lewis, D. [1973] 1986b. "Causation." Reprint with postscripts, *Philosophical Papers*, vol. 2, 159–213. Oxford: Oxford University Press.
- Lewis, D. [1979] 1986c. "Counterfactual Dependence and Time's Arrow." Reprint with postscripts, *Philosophical Papers*, vol. 2, 32–66. Oxford: Oxford University Press.
- Lewis, D. 1986d. *Philosophical Papers*. Vol. 2. Oxford: Oxford University Press.
- Lewis, D. 1989. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society* 63 (Supplement): 113–37.
- Lewis, D. 2000. "Causation as Influence." *Journal of Philosophy* 97: 182–97.
- Lucas, R. 1983. "Econometric Policy Evaluation: A Critique." In *Carnegie-Rochester Conference on Public Policy: Supplementary Series to the Journal of Monetary Economics*, 1, ed. K. Brunner and A. Meltzer, 257–84. Amsterdam: North Holland.
- Luce, R. D., and H. Raiffa. 1957. *Games and Decisions*. New York: Wiley.

- Lyon, A. 1977. "The Immutable Laws of Naure." In *Proceedings of the Aristotelian Society 1976–77*, 107–26. London: Compton Press.
- Mackie, J. 1974. *The Cement of the Universe*. Oxford: Oxford University Press.
- McDermott, M. 1995. "Redundant Causation." *British Journal for the Philosophy of Science* 46: 523–44.
- Meek, C., and C. Glymour. 1994. "Conditioning and Intervening." *British Journal for the Philosophy of Science* 45: 1001–21.
- Mellor, D. 1995. *The Facts of Causation*. London: Routledge.
- Menzies, P., and H. Price. 1993 "Causation as a Secondary Quality." *British Journal for the Philosophy of Science* 44: 187–203.
- Mitchell, S. 1997. "Pragmatic Laws." *PSA 96* (Supplement to *Philosophy of Science* 64.4): S468–S479.
- Mitchell, S. 2000. "Dimensions of Scientific Law." *Philosophy of Science* 67: 242–65.
- Morrison, M. 2000. *Unifying Scientific Theories*. Cambridge, UK: Cambridge University Press.
- Morton, A. 1973. "If I Were a Dry Well-Made Match." *Dialogue* 12: 322–24.
- Nagel, E. 1961. *The Structure of Science*. New York: Harcourt, Brace, and World.
- Neander, K. 1988. "What Does Natural Selection Explain? Correction to Sober." *Philosophy of Science* 55: 422–26.
- Nerlich, G. 1979. "What Can Geometry Explain?" *British Journal for the Philosophy of Science* 30: 69–83.
- Newbold, P., and T. Bos. 1985. *Stochastic Parameter Regression Models*. Beverly Hills: Sage.
- Newton, R. 1993. *What Makes Nature Tick?* Cambridge, MA: Harvard University Press.
- Oakeshott, M. 1966. "Historical Continuity and Causal Analysis." In *Philosophical Analysis and History*, ed. W. Dray, 192–212. New York: Harper and Row.
- Ohanian, H. 1976. *Gravitation and Spacetime*. New York: Norton.
- Orcutt, G. 1952. "Actions, Consequences, and Causal Relations." *Review of Economics and Statistics* 34: 305–13.
- Papineau, D. 1991. "Correlations and Causes." *British Journal for the Philosophy of Science* 42: 397–412.
- Pearl, J. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82: 669–88.
- Pearl, J. 2000a. *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge University Press.
- Pearl, J. 2000b. "Direct and Indirect Effects." Technical report, R-273. Department of Computer Science, University of California, Los Angeles.
- Pearl, J. 2001. "Causality." [www.cs.ucla.edu/~judeal](http://www.cs.ucla.edu/~judeal). Accessed August 1, 2001.
- Phillips, A. W. 1958. "The Relation between Unemployment and the Rate of Change of Money Wages in the United Kingdom, 1861–1957." *Economica* NS 25.2: 283–99.
- Pietroski, P., and G. Rey. 1995. "When Other Things Aren't Equal: Saving *Ceteris Paribus* Laws from Vacuity." *British Journal for the Philosophy of Science* 46: 81–110.
- Planck, A. [1960] 1997. *Survey of Physical Theory*. Translated by R. Jones and D. H. Williams. New York: Dover. Originally published as *Physikalische Rundblicke* (Leipzig).
- Plotkin, H. 1997. *Evolution in Mind*. London: Penguin.
- Popper, K. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.
- Price, H. 1991. "Agency and Probabalistic Causality." *British Journal for the Philosophy of Science* 42: 157–76.

- Railton, P. 1978. "A Deductive-Nomological Model of Probabilistic Explanation." *Philosophy of Science* 45: 206–26.
- Railton, P. 1981. "Probability, Explanation, and Information." *Synthese* 48: 233–56.
- Redhead, M. 1987. *Incompleteness, Nonlocality, and Realism: A Prolegomenon to the Philosophy of Quantum Mechanics*. Oxford: Oxford University Press.
- Reichenbach, H. 1956. *The Direction of Time*. Berkeley: University of California Press.
- Rosenberg, A. 1994. *Instrumental Biology or the Disunity of Science*. Chicago: University of Chicago Press.
- Ruben, D. 1990. *Explaining Explanation*. London: Routledge.
- Rubin, D. 1986. "Comment: Which Ifs Have Causal Answers?" *Journal of the American Statistical Association* 81: 961–62.
- Salmon, W. 1971. "Statistical Explanation." In *Statistical Explanation and Statistical Relevance*, ed. W. Salmon, 29–87. Pittsburgh: University of Pittsburgh Press.
- Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
- Salmon, W. 1989. *Four Decades of Scientific Explanation*. Minneapolis: University of Minnesota Press.
- Salmon, W. 1994. "Causality without Counterfactuals." *Philosophy of Science* 61: 297–312.
- Salmon, W. 1997. "Causality and Explanation: A Reply to Two Critiques." *Philosophy of Science* 64: 461–77.
- Salmon, W., and P. Kitcher, eds. 1989. *Minnesota Studies in the Philosophy of Science*, Vol. 13: *Scientific Explanation*. Minneapolis: University of Minnesota Press.
- Satz, D., and J. Ferejohn. 1994. "Rational Choice and Social Theory." *Journal of Philosophy* 91: 71–87.
- Schaffer, J. 2000a. "Causation by Disconnection." *Philosophy of Science* 67: 285–300.
- Schaffer, J. 2000b. "Trumping Preemption." *Journal of Philosophy* 97: 165–181.
- Scriven, M. 1959a. "Explanation and Prediction in Evolutionary Theory." *Science* 30: 477–82.
- Scriven, M. 1959b. "Truisms as the Grounds of Historical Explanations." In *The Nature of Historical Explanation*, 443–75, ed. P. Gardiner. Oxford: Oxford University Press.
- Scriven, M. 1962. "Explanations, Predictions, and Laws." in *Minnesota Studies in the Philosophy of Science*, Vol. 3: *Scientific Explanation, Space, and Time*, ed. H. Feigl and G. Maxwell, 170–230. Minneapolis: University of Minnesota Press.
- Skyrms, B. 1980. *Causal Necessity*. New Haven: Yale University Press.
- Skyrms, B. 1984. "EPR: Lessons for Metaphysics." *Midwest Studies in Philosophy* 9: 245–55.
- Skyrms, B., and K. Lambert. 1995. "The Middle Ground: Resiliency and Laws in the Web of Belief." In *Laws of Nature: Essays on the Philosophical, Scientific, and Historical Dimensions*, ed. F. Weinert. Berlin: Walter de Gruyter.
- Smith, C., and N. Wise. 1989. *Industry and Empire: A Biographical Study of Lord Kelvin*. Cambridge, UK: Cambridge University Press.
- Sober, E. 1983. "Equilibrium Explanation." *Philosophical Studies* 43: 201–10.
- Sober, E. 1984. *The Nature of Selection*. Cambridge, MA: MIT Press.
- Sober, E. 1985. "Two Concepts of Cause." In *PSA 1984*, vol. 2, ed. P. Asquith and P. Kitcher, 405–24. East Lansing, MI: Philosophy of Science Association.
- Sober, E. 1988. *Reconstructing the Past: Parsimony, Evolution and Inference*. Cambridge, MA: MIT Press.

- Sober, E. 1994. "Temporally Oriented Laws." In E. Sober, *From a Biological Point of View*, 233–51. Cambridge: Cambridge University Press.
- Sober, E. 1995. "Natural Selection and Distributive Explanation: A Reply to Neander." *British Journal for the Philosophy of Science* 46: 384–97.
- Sober, E. 1997. "Two Outbreaks of Lawlessness in Recent Philosophy of Biology." In *PSA96*, ed. L. Darden, S458–S467. East Lansing, MI: Philosophy of Science Association.
- Sosa, E., and M. Tooley, eds. 1993. *Causation*. Oxford: Oxford University Press.
- Sperber, D., D. Premack, and A. Premack. 1995. *Causal Cognition*. Oxford: Oxford University Press.
- Spirtes, P., C. Glymour, and R. Scheines. [1993] 2000. *Causation, Prediction and Search*. 1st ed. New York: Springer-Verlag.
- Squire, L., and E. Kandel. 1999. *Memory: From Mind to Molecules*. New York: Scientific American.
- Stafford, F. 1985. "Income-Maintenance Policy and Work Effort: Learning from Experiments and Labor Market Studies." In *Social Experimentation*, ed. J. Hausman and D. Wise. Chicago: University of Chicago Press.
- Steiner, M. 1978. "Mathematical Explanation." *Philosophical Studies* 34: 135–51.
- Strotz, R. H., and H. O. A. Wold. 1960. "Recursive vs. Non-recursive Systems: An Attempt at a Synthesis." *Econometrica* 28: 417–27.
- Tufte, E. 1974. *Data Analysis for Politics and Policy*. Englewood Cliffs, NJ: Prentice-Hall.
- van Fraassen, B.C. 1980. *The Scientific Image*. Oxford: Oxford University Press.
- van Fraassen, B. C. 1989. *Laws and Symmetry*. Oxford: Clarendon Press.
- Veblen, E. 1975. *The Manchester Union-Leader in New Hampshire Elections*. Hanover, NH: University Press of New England.
- von Wright, G. 1971. *Explanation and Understanding*. Ithaca, NY: Cornell University Press.
- Waters, C. K. 1998. "Causal Regularities in the Biological World of Contingent Distributions." *Biology and Philosophy* 13: 5–36.
- Weinberg, R. 1985. "The Molecules of Life." *Scientific American* 253.4: 48–57.
- Wigner, E. 1967. *Symmetries and Reflections: Scientific Essays*. Bloomington: Indiana University Press, 1967.
- Woodward, J. 1984. "Explanatory Asymmetries." *Philosophy of Science* 51: 421–42.
- Woodward, J. 1986. "Are Singular Causal Explanations Implicit Covering-Law Explanations?" *Canadian Journal of Philosophy* (June): 253–79.
- Woodward, J. 1989. "The Causal/Mechanical Model of Explanation." In *Minnesota Studies in the Philosophy of Science*, Vol. 13: *Scientific Explanation*, ed. W. Salmon and P. Kitcher, 357–83. Minneapolis: University of Minnesota Press.
- Woodward, J. 1990. "Supervenience and Singular Causal Claims." In *Explanation and Its Limits* (Proceedings of the Royal Institute of Philosophy Conference), ed. Dudley Knowles, 211–46. Cambridge, UK: Cambridge University Press.
- Woodward, J. 1993a. "Capacities and Invariance." In *Philosophical Problems of the Internal and External Worlds: Essays Concerning the Philosophy of Adolph Grünbaum*, ed. J. Earman, A. Janis, G. Massey, and N. Rescher, 283–328. Pittsburgh: University of Pittsburgh Press.
- Woodward, J. [1984] 1993b. "A Theory of Singular Causal Explanation." Reprint, *Explanation*, ed. D. Reuben, 246–74. Oxford: Oxford University Press.

- Woodward, J. 1995. "Causality and Explanation in Econometrics." In *On the Reliability of Economic Models: Essays in the Philosophy of Economics*, ed. D. Little, 9–61. Dordrecht: Kluwer.
- Woodward, J. 1997a. "Causal Modeling, Probabilities and Invariance." In *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*, ed. V. McKim and S. Turner, 265–317. Notre Dame, IN: University of Notre Dame Press.
- Woodward, J. 1997b. "Explanation, Invariance and Intervention." *PSA 1996 2*: S26–S41.
- Woodward, J. 1998. "Causal Independence and Faithfulness." *Multivariate Behavioral Research* 33: 129–48.
- Woodward, J. 1999. "Causal Interpretation in Systems of Equations." *Synthese* 121: 199–257.
- Woodward, J. 2000. "Explanation and Invariance in the Special Sciences." *British Journal for the Philosophy of Science* 51: 197–254.
- Woodward, J. 2001a. "Causation and Manipulability." *Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu>. Accessed November 2001.
- Woodward, J. 2001b. "Law and Explanation in Biology: Invariance Is the Kind of Stability That Matters." *Philosophy of Science* 68: 1–20.
- Woodward, J. 2001c. "Probabilistic Causality, Direct Causes, and Counterfactual Dependence." In *Stochastic Causality*, ed. M. Galavotti, P. Suppes, and D. Costantini, 39–63. Stanford: CSLI.
- Woodward, J. 2003. "Experimentation, Causal Inference, and Instrumental Realism." In *The Philosophy of Scientific Experimentation*, ed. H. Radder, 87–118. Pittsburgh: University of Pittsburgh Press.
- Woodward, J. Forthcoming a. "A Plea for Invariance." Unpublished ms. Paper read at Conference on "Explanation in the Natural and Social Sciences." Ghent, Belgium, 2002.
- Woodward, J. 2002a. "There Is No Such Thing as a Ceteris Paribus Law." *Erkenntnis* 57: 303–328.
- Woodward, J. 2002b. "What Is a Mechanism? A Counterfactual Account." In *PSA 2000*, Part 2. *Philosophy of Science* 69: S366–77.
- Woodward, J., and C. Hitchcock. 2003. "Explanatory Generalizations, Part 1: A Counterfactual Account" *Nous* 37: 1–24.
- Woodward, J. Forthcoming b. "Invariance, Modularity, and All That: Comments on Cartwright." Unpublished ms. Paper read at a conference on Nancy Cartwright's Philosophy of Science." Konstanz, Germany. December 2002.
- Worrall, J. 1984. "An Unreal Image." *British Journal for the Philosophy of Science* 35: 65–80.

## Index

- actual cause, 40, 74–86  
agency theories of causation, 23–27,  
123–27  
Anscombe, G., 72  
Armstrong, D., 388  
Ayer, A., 394
- Barkan, D., 12  
Barnes, E., 361–62  
Beatty, J., 276, 302–7  
Bennett, J., 67, 135  
Bogen, J., 7  
Brandon, R., 394
- Campbell, D., 25, 69  
Carroll, J., 268  
Cartwright, N., 32–46, 61–63, 106,  
108, 196, 266, 380, 381, 385,  
396  
Causal Markov Condition, 64, 106–7,  
378, 385  
causal mechanical model of explanation,  
350–58  
causal overdetermination, 82–84  
causal pre-emption, 77–80  
causal relata, 111–14  
causation  
as a cluster concept, 91–93  
conserved quantity theory of, 30–31,  
350–58  
contributing cause, 50  
and counterfactuals, 88–91, 122–23,  
356–58  
versus description, 318–21  
direct, 42, 51–55  
and hypothetical experiments, 114–  
17  
and instrumental conditioning, 34  
and laws, 146–47, 168–74  
and omission, 86–91  
practical theories of, 30–31  
probabilistic theories of, 61–65  
realism about, 118–19  
reductionist accounts of, 20–22,  
104–7  
regularity theory, 31–32  
and representation sensitivity,  
55–57  
and reproducibility, 41–42, 70–74  
and serious possibility, 80–81, 86–91  
and spatio-temporal continuity,  
147–49  
total (TC), 50, 51  
transitivity of, 57–59  
and unanimity condition, 62, 149  
ceteris paribus laws, 307–11  
Churchland, P., 5  
Collingwood, R., 12, 25  
Collins, J., 87  
contrastive focus, 67, 145–46  
Cook, T., 25, 69  
Cooley, T., 396  
Cooper, G., 394  
counterfactuals  
and causation, 88–91, 122–23,  
356–58  
and explanation, 187–94, 196–203  
and invariance, 279–85  
Cowie, F., 376
- Davidson, D., 163–64, 165, 388  
Deductive-Nomological Model of  
Explanation, 152–54, 184–86  
counterexamples to, 154–55



- Devitt, M., 389  
 direct generalization, 165  
 directed graphs, 38–39, 42  
 Dowe, P., 7, 30, 35  
 Ducasse, C., 72  
 Dummett, M., 11  
 Dunn, M., 276  
 Dupre, J., 381
- Earman, J., 294–95  
 Eells, E., 380  
 epistemology and unification, 369–71  
 epistemological constraints on  
 explanation, 22, 179–81  
 epistemological thesis, 174–75  
 equations as devices for representing  
 causal relationships, 42–45  
 error term, 325–27  
 Ervin, E., 29  
 exception incorporation, 273  
 explanation  
 in biology, 7–9, 302–7  
 and causation, 5–7, 182–84  
 and change, 208–9, 233–36  
 counterfactual theory of, 187–94,  
 196–203  
 and derivation, 185, 200–202  
 desiderata for a theory of, 23–24  
 versus description, 5  
 and invariance, 239–42  
 and laws, 182–84, 236–38  
 in mathematics, 220–21  
 minimal covering law, 194  
 nomothetic, 160, 181, 203–4  
 and omissions, 224–26  
 and ontology, 223–24  
 and pragmatics, 226–33  
 thesis, 164, 175–81
- faithfulness, 49, 64–65  
 Fiorina, M., 276, 278  
 Fodor, J., 191, 307, 309, 392, 393  
 Freedman, D., 25, 319–21, 395  
 Friedman, M., 155, 358, 397  
 Frisch, R., 321
- Garcia, J., 29  
 Gardiner, P., 387  
 Garfinkel, A., 392  
 Gasking, D., 12, 25
- Glymour, C., 34, 38, 39, 46, 64,  
 376, 377, 380, 382, 385, 395,  
 397  
 Green, D., 276, 278
- Haavelmo T., 18, 39, 258–60, 310–  
 11, 321, 333  
 Hajek, A., 381, 387, 392  
 Hall, N., 379, 381, 391–92  
 Halpern, J., 382  
 Hausman, D., 37, 46, 48, 64, 108,  
 135, 255–56, 286, 307, 377,  
 385, 392, 394–95  
 Healey, R., 92, 256–57, 391, 393  
 Heathcote, A., 388  
 Hempel, C., 152–54, 161, 164–65,  
 170–71, 175, 182, 194–95, 197–  
 99, 376, 393  
 Hidden Structure Strategy, 83, 157–  
 61, 175–81  
 Hitchcock, C., 351–53, 376, 378, 379,  
 382, 384, 385, 386, 393  
 Holland, P., 25, 117, 386  
 Hoover, K., 25, 393  
 Horgan, T., 388  
 Horwich, P., 137  
 Hughes, R., 376  
 Humphreys, P., 12, 196
- ideal explanatory text, 176–79  
 independent specification, 273  
 internal validity, 69  
 interventions, 14, 15, 46–48,  
 94–103  
 and anthropomorphism, 103–4  
 and circularity, 104–7  
 defined, 98  
 other notions of, 107–11  
 and possibility, 127–33  
 invariance, 15, 69–70, 239–54  
 in biology, 302–37  
 and counterfactuals, 279–85  
 degrees of, 257–65  
 in economics, 263–64  
 and exceptionlessness, 270–73  
 and laws, 15–16, 206–8, 265–68,  
 285–88  
 and possible cause generalizations,  
 311–14  
 and probabilistic causation, 339–42

- and qualitative predicates, 268–69
- and resiliency, 299–302
- and scope, 269–70
- and symmetry, 262–63
- Irzik, G., 396
- Jones, M., 108–10, 385
- Keller, E., 384
- Keeble, S., 29
- Kim, J., 171
- Kincaid, H., 286, 392
- Kitcher, P., 147, 155, 159, 358–73
- Koelling, R., 29
- Kvart, I., 387, 391
- Lambert, K., 299–302
- laws
  - criteria for, 182–84, 265–302
  - and invariance, 235–40
  - role in explanation, 182–84, 206–8, 236–38
  - weak and strong notions, 166–67
- LeRoy, S., 396
- Leslie, A., 29
- Lewis, D., 3, 74, 81, 133–45, 159, 288, 376, 379, 382, 386, 387
- Mackie, J., 5, 67
- manipulability account of causation, 9, 25–28, 32–38, 59
- McDermott, M., 57, 384
- meaning thesis, 163, 168–73
- mechanisms, 48–49, 52–53
- Meek, C., 377
- Mellor, D., 376
- Menzies, P., 104, 118, 123–27, 386
- Mill-Ramsey-Lewis theory of law, 288–95
- Mitchell, S., 295–99, 394
- modularity, 48–49, 327–39
- Morrison, M., 365–66, 397
- Morton, A., 283
- Nagel, E., 268
- NC**, 45
- NC\***, 57
- Nerlich, G., 6
- Newton, R., 26
- Oakeshott, M., 313–15
- Oppenheim, P., 152, 393
- Orcutt, G., 26
- Papineau, D., 395, 396
- Pearl, J., v, 4, 38, 46–48, 64, 83, 107–8, 110–11, 335, 377, 378, 380, 382, 386, 396
- Phillips curve, 263–64
- Pietroski, P., 271, 307, 392, 393, 394
- Popper, K., 387
- possible cause generalizations, 214–15, 217, 342–46
- Price, H., 104, 118, 123–27, 386
- Railton, P., 159, 175–79
- rational choice models, 276–79
- Redhead, M., 256
- regression models, 315–18
  - and causal interpretation, 321–27
  - role of error term in, 325–27
- Reichenbach, H., 64
- Rey, G., 271, 307, 309, 392, 393, 394
- Ruben, D., 3
- Rubin, D., 386
- Salmon, W., 30, 35, 133, 145–46, 154–55, 182, 200–201, 350–58, 376, 397
- SC**, 45
- Schaffer, J., 81, 382, 383
- Scheines, R., 38, 39, 46, 64, 377, 380, 382, 385
- Scriven, M., 155–56, 161, 171, 389, 390
- Shapiro, I., 276, 278
- singular causal explanation, 209–20
- Skyrms, B., 91–93, 299–302, 387, 392, 393, 394
- Smith, C., 12
- Sober, E., 6, 62, 75, 394
- Sosa, E., 377
- Spirtes, P., 38, 39, 46, 64, 377, 380, 382, 385
- Sterelny, K., 389
- structural equations, 327–36
  - inputs to, 342–46
  - as population specific, 346–49

- Tooley, M., 377  
trumping, 81–82  
type-causation, 40
- underlying thesis, 163, 173–74  
unificationist models of explanation,  
358–73
- van Fraassen, B., 36, 266, 292, 336,  
392, 393
- variables, 39–40, 45  
von Wright, G., 12, 104
- Waters, K., 297  
Weinberg, R., 9  
what-if-things-had-been-different  
question, 11, 187–94,  
196–203  
Wise, N., 12  
Worrall, J., 145