EDITED BY

HELEN
BEEBEE

CHRISTOPHER
HITCHCOCK

PETER
MENZIES

The Oxford Handbook *of*
CAUSATION

THE OXFORD HANDBOOK OF

CAUSATION

THE OXFORD HANDBOOK OF

CAUSATION

Edited by

HELEN BEEBEE,

CHRISTOPHER HITCHCOCK,
AND
PETER MENZIES

ACKNOWLEDGEMENTS

CONTENTS

LIST OF CONTRIBUTORS

Helen Beebee is Professor of Philosophy at the University of Birmingham. Her publications include 'Seeing Causing' in *Proceedings of the Aristotelian Society* 103 (2003); 'Causing and Nothingness' in J. Collins, E. J. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals* (MIT, 2004); and *Hume on Causation* (Routledge, 2006).

Sarah Broadie is Wardlaw Professor of Philosophy at the University of St Andrews. As Sarah Waterlow she has published *Nature, Change, and Agency in Aristotle's Physics* and *Passage and Possibility* (both Oxford University Press, 1982); as Sarah Broadie *Ethics with Aristotle* (Oxford University Press, 1991) and *Aristotle,* Nicomachean Ethics (with Christopher Rowe). Her collection *Aristotle and Beyond: Essays on Metaphysics and Ethics* (Cambridge University Press, 2007) includes several pieces on causation.

John W. Carroll is Professor of Philosophy at North Carolina State University. His publications include 'Property-Level Causation?' in *Philosophical Studies* 63 (1991) and *Laws of Nature* (Cambridge University Press, 1994).

Kenneth Clatterbaugh is Professor of Philosophy and Chair of the Department of Philosophy at the University of Washington, Seattle. His publications include 'Descartes's Causal Likeness Principle' in *Philosophical Review* 89 (1980); 'Cartesian Causality, Explanation, and Divine Concurrence' in *History of Philosophy Quarterly* 12 (1995); and *The Causation Debate in Modern Philosophy 1637-1739* (Routledge, 1999).

Helen Daly is a graduate student in philosophy at the University of Arizona.

David Danks is Associate Professor of Philosophy at Carnegie Mellon University and Research Scientist at the Florida Institute for Human and Machine Cognition. His publications include 'Causal Learning from Observations and Manipulations' in M. Lovett and P. Shah (eds.), *Thinking with Data* (Lawrence Erlbaum, 2007); and 'The Supposed Competition between Theories of Human Causal Inference' in *Philosophical Psychology* 18 (2005).

Phil Dowe is Reader in Philosophy at the University of Queensland. His publications include *Physical Causation* (Cambridge University Press, 2000) and 'Every Now and Then: A-theory and Loops in Time' in *Journal of Philosophy* (2009).

Douglas Ehring is the William Edward Easterwood Professor of Philosophy at Southern Methodist University in Dallas, Texas. His publications include 'Causal Relata' in *Synthese* 73 (1987) and *Causation and Persistence: A Theory of Causation* (Oxford University Press, 1997).

Don Garrett is Professor of Philosophy at New York University. His publications include 'The Representation of Causation and Hume's Two Definitions of "Cause"' in *Noûs* 27 (1993); *Cognition and Commitment in Hume's Philosophy* (Oxford University Press, 1997); and 'Hume's Naturalistic Theory of Representation' in *Synthese* 152 (2006).

Stuart Glennan is Professor of Philosophy at Butler University in Indianapolis. His publications include 'Mechanisms and the Nature of Causation' in *Erkenntnis* 44 (1996) and 'Modeling Mechanisms' in *Studies in the History and Philosophy of the Biological and Biomedical Sciences* 36 (2005).

Clark Glymour is Alumni University Professor at Carnegie Mellon University and Senior Research Scientist at the Florida Institute for Human and Machine Cognition. He has been a Guggenheim Fellow, a Fellow of the Center for Advanced Study in the Social Sciences, and is a Fellow of the Statistics Section of the American Association for the Advancement of Science.

Peter Godfrey-Smith is Professor of Philosophy at Harvard University. His publications include *Theory and Reality: An Introduction to the Philosophy of Science* (Chicago University Press, 2003) and *Darwinian Populations and Natural Selection* (Oxford University Press, 2009).

Richard Healey is Professor of Philosophy at the University of Arizona. His publications include *The Philosophy of Quantum Mechanics* (Cambridge University Press, 1989); 'Chasing Quantum Causes: How Wild is the Goose?' in *Philosophical Topics* 20 (1992); and 'Nonseparable Processes and Causal Explanation' in *Studies in History and Philosophy of Science* 25 (1994).

Christopher Hitchcock is Professor of Philosophy at the California Institute of Technology. He has published numerous papers in the Philosophy of Science, especially on the topic of causation, in journals such as *Journal of Philosophy, Philosophical Review, Noûs, Philosophy of Science,* and *British Journal for the Philosophy of Science.*

Carl Hoefer is ICREA Research Professor in Philosophy at the Autonomous University of Barcelona. His publications include 'Humean Effective Strategies' in P. Hajek, L. Valdes-Villanueva, and D. Westerstahl (eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 12th International Congress* (King's College London, 2005) and 'Causality and Determinism: Tension, or Outright Conflict?' in *Revista de Filosofía* 29 (2004).

Terry Horgan is Professor of Philosophy at the University of Arizona. His publications include 'Mental Quausation' in *Philosophical Perspectives* 3 (1989); 'Kim on Mental Causation and Causal Exclusion' in *Philosophical Perspectives* 11 (1998); 'Causal Compatibilism and the Exclusion Problem' in *Theoria* 16 (2001); and 'Mental Causation and the Agent-Exclusion Problem' in *Erkenntnis* 67 (2007).

Paul Humphreys is Professor of Philosophy at the University of Virginia. He is the author of *The Chances of Explanation* (Oxford University Press, 1989) and 'Causation' in W. H. Newton-Smith (ed.), *A Companion to the Philosophy of Science* (Blackwell, 1999).

Frank Jackson teaches at Princeton University each autumn and is at La Trobe University or the Australian National University the rest of the year. His publications include 'Mental Causation: the State of the Art' in *Mind* 105 (1996); 'Causation in the Philosophy of Mind' (with Philip Pettit) in *Philosophy and Phenomenological Research* 50 Supplementary (1990); and *The Philosophy of Mind and Cognition* (with David Braddon-Mitchell) (Basil Blackwell, 2nd edn., 2007).

Harold Kincaid is Professor and Chair of the Department of Philosophy at the University of Alabama at Birmingham. He has written multiple books and numerous articles on issues in the philosophy of the social sciences.

Marc Lange is Bowman and Gordon Gray Distinguished Professor of Philosophy at the University of North Carolina at Chapel Hill. His publications include *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass* (Black-well, 2002) and 'How Can Instantaneous Velocity Fulfill its Causal Role?' in *Philosophical Review* 114 (2005).

Peter Lipton passed away on 25 November 2007, as this volume was being prepared. He was the author of *Inference to the Best Explanation* (Routledge, 1991) and numerous articles in the philosophy of science. He was Head of Cambridge University's Department of History and Philosophy of Science for many years. He earned a reputation as a gifted teacher and caring mentor. He will be sorely missed by family and friends, students and colleagues, and the profession of philosophy.

John Marenbon is Senior Research Fellow at Trinity College, Cambridge. His publications include *Abelard* (Cambridge University Press, 1987); *Boethius* (Oxford University Press, 2003) and *Medieval Philosophy: An Historical and Philosophical Introduction* (Routledge, 2007).

Cei Maslen is a lecturer in Philosophy at Victoria University of Wellington, New Zealand. Her publications include 'Causes, Contrasts and the Nontransitivity of Causation' in J. Collins, E. J. Hall, and L.A. Paul (eds.), *Causation and Counter-factuals* (MIT, 2004) and 'Counterfactuals as Short Stories' (with Seahwa Kim) in *Philosophical Studies* 129 (2006).

Alfred R. Mele is the William H. and Lucyle T. Werkmeister Professor of Philosophy at Florida State University. His publications include *Motivation and Agency* (Oxford University Press, 2003); *Free Will and Luck* (Oxford University Press, 2006); and *Effective Intentions: the Power of Conscious Will* (Oxford University Press, 2009).

Peter Menzies is Professor of Philosophy at Macquarie University in Sydney. He has written numerous articles on causation, including most recently 'Causation in Context' in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited* (Oxford University Press, 2007) and 'Causal Models, Token Causation, and Processes' in *Philosophy of Science* 71 (2004).

Stephen Mumford is Professor of Metaphysics at the University of Nottingham. His books include *Dispositions* (Oxford University Press, 1998); *Laws in Nature* (Routledge, 2004); and *David Armstrong* (Acumen, 2007). He also edited George Molnar's posthumous book *Powers* (Oxford University Press, 2003).

Ram Neta is Associate Professor of Philosophy at the University of North Carolina, Chapel Hill. His publications include 'S knows that *p*' in *Noûs* 36 (2002); and 'Contextualism and a Puzzle about Seeing' in *Philosophical Studies* 136 (2007).

Samir Okasha is Professor of Philosophy of Science at the University of Bristol. He is the author of numerous papers in Philosophy of Science and Philosophy of Biology. His book *Evolution and the Levels of Selection* was published by Oxford University Press in 2006.

L. A. Paul is Associate Professor of Philosophy at University of North Carolina at Chapel Hill. Her publications include 'Aspect Causation' in *Journal of Philosophy* 97 (2000); the co-edited volume (with John Collins and Ned Hall) *Causation and Counterfactuals* (MIT, 2004); and *Causation: A User's Guide,* co-authored with Ned Hall (forthcoming from Oxford University Press).

Huw Price is ARC Federation Fellow and Challis Professor of Philosophy at the University of Sydney. His publications include *Facts and the Function of Truth* (Black-well, 1988); and *Time's Arrow and Archimedes' Point* (Oxford University Press, 1996). He is also co-editor (with Richard Corry) of *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited* (Oxford University Press, 2007).

Stathis Psillos is Professor of Philosophy of Science at the University of Athens. He is the author of *Scientific Realism: How Science Tracks Truth* (Routledge, 1999); *Causation and Explanation* (Acumen and McGill-Queens University Press, 2002); and *Philosophy of Science A-Z* (Edinburgh University Press, 2007). He edited with Martin Curd *The Routledge Companion to Philosophy of Science* (Routledge, 2008).

Carolina Sartorio is Associate Professor of Philosophy at the University of Arizona. Her publications include 'How to be Responsible for Something without Causing It' in *Philosophical Perspectives* 18 (2004); and 'Moral Inertia' in *Philosophical Studies* 140 (2008).

Lawrence Sklar is Distinguished University Professor at the University of Michigan. His publications include *Physics and Chance* (Cambridge University Press, 1993); *Theory and Truth* (Oxford University Press, 2000); and *Philosophy of Physics* (Oxford University Press, 1995).

Jane Stapleton is Ernest E. Smith Professor of Law at the University of Texas School of Law, Professor at the Australian National University College of Law, and Statutory Visiting Professor at Oxford University Faculty of Law. Her publications include 'Choosing What We Mean by "Causation" in the Law', *Missouri Law Review* 73 (2008).

Michael Stöltzner is Associate Professor of Philosophy at the University of South Carolina at Columbia. His publications include 'Vienna Indeterminism: Mach, Boltzmann, Exner' in *Synthese* 119 (1999) and 'The Least Action Principle as the Logical Empiricist's Shibboleth' in *Studies in History and Philosophy of Modern Physics* 34 (2003).

Michael Tooley is Distinguished College Professor in Philosophy at the University of Colorado at Boulder. His publications include *Causation: A Realist Approach* (Oxford University Press, 1987); *Time, Tense, and Causation* (Oxford University Press, 1997); 'Causation: Reductionism Versus Realism' in *Philosophy and Phenomenological Research,* 50, Supplement (1990); 'The Nature of Causation: A Singularist Account' in D. Copp (ed.), *Canadian Philosophers: Celebrating Twenty Years of the Canadian Journal of Philosophy,*

Canadian Journal of Philosophy, Supplementary 16 (1990); and 'Causation and Supervenience' in M. Loux and D. Zimmerman (eds.), *Oxford Handbook of Metaphysics* (Oxford University Press, 2003).

Eric Watkins is Professor of Philosophy at the University of California, San Diego. His publications include *Kant and the Metaphysics of Causality* (Cambridge University Press, 2005) and *Kant and the Sciences* (Oxford University Press, 2000).

Brad Weslake is Assistant Professor of Philosophy at the University of Rochester.

Jon Williamson is Professor of Reasoning, Inference and Scientific Method in the Philosophy Department at the University of Kent, UK. His publications include *Bayesian Nets and Causality* (Oxford, 2005) and 'Causality', *Handbook of Philosophical Logic* 14 (2007).

James F. Woodward is J. O. and Juliette Koepfli Professor of Humanities and Professor of Philosophy at the California Institute of Technology. His publications include *Making Things Happen: A Theory of Causal Explanation* (Oxford University Press, 2003) and 'Sensitive and Insensitive Causation' in *Philosophical Review* 115 (2006).

INTRODUCTION
HELEN BEEBEE CHRISTOPHER HITCHCOCK PETER MENZIES

PHILOSOPHERS have been interested in the nature of causation for as long as there has been philosophy. They have been interested in what we say about the world when we say that one thing caused another, and in whether there is anything in the world that answers to the causal claims we make about it. Despite the attention, there is still very little agreement on the most central question concerning causation: what *is* it? Is it a matter of the instantiation of regularities or laws, or counterfactual dependence, or manipulability, or transfer of energy, for example? One reason for the lack of a consensus view is the sheer difficulty of the task; anyone familiar with the causation debate as it has been conducted in recent years will be familiar with a vast range of theories and counterexamples, which collectively can lead one to suspect that no univocal analysis of the concept of causation is possible.

Another, related reason for the lack of consensus is that different authors have radically different views about the metaphysical status of causation. Most of us agree that colours are not a part of the fundamental fabric of the universe, while, say, subatomic particles probably are. When it comes to causation, however, there is no such consensus. Some authors think that causation is a relatively non-fundamental feature of the world, a feature that can be understood in terms of other, more fundamental features, such as regularities. Some think that in some sense causation is not a feature of reality at all, so that the concept of 'cause' is something that we project or impose upon the goings-on in the world around us. Still others think that causation is about as fundamental as it gets, so that the notion of a fundamental layer of reality that can in principle be described without deploying causal terms is incoherent, or at least obviously mistaken.

A third reason is the fact that philosophical theories of causation are hostage to developments in the sciences in a way that many other philosophical theories are not. For example, Newton's celestial mechanics seems to posit instantaneous action at a distance, and quantum mechanics seems to tell us that the fundamental processes of our world are indeterministic. Both developments challenged existing assumptions about how causes could operate in our world. Moreover, many philosophical theories of causation draw upon the resources of mathematical or scientific theories in their formulation. Probabilistic theories deploy the resources of probability theory, causal process theories borrow concepts from the special theory of relativity, and so on.

A fourth reason is that the concept of causation is used in many different contexts, in ways that are not obviously compatible with one another. The physicist's conception of causation, for example, may seem to be utterly different from the lawyer's conception, which may be utterly different again from the ordinary notion we deploy when dealing with the situations

that confront us in our daily lives.

A fifth reason is the centrality of the concept of causation to other areas of philosophy. Philosophical theories that appeal to causation span the philosophical spectrum: they are to be found in ethics, in epistemology, and in virtually every topic in metaphysics: free will, personal identity, time, universals, laws of nature, and so on. One's choice of theory of causation can have radical consequences for these other areas. A theory of causation that builds temporal priority into the definition of cause rules out the possibility of a causal analysis of the direction of time; a theory that rules out causation by omission sits uneasily with consequentialism in ethics; a 'Humean' analysis of causation is incompatible with the view that all fundamental universals are dispositional; and so on.

The chapters in this book explore these and other issues in some detail, while, for the most part, starting at a sufficiently introductory level that they should be wholly or mostly understandable to an upper-level undergraduate. The chapters aim, in general, to survey the existing literature and give a sense of the major points of controversy, while at the same time maintaining a distinctive authorial voice: positive, and often controversial, positions are sketched and defended, so that the chapters will be interesting not just to the novice, but to the philosopher who is already well acquainted with the ongoing debate.

The book is divided into seven parts. We provide a very brief overview of each part, along with a short abstract of each chapter.

PART I: THE HISTORY OF CAUSATION

Part I offers some historical background to many of the issues and disputes surrounding causation that concern philosophers today. Causation's role in both explanation and inference, for example, has been recognized since the ancient Greeks. And, while some issues that preoccupied some of our predecessors have understandably lost the prominence they once had (the causal role of God in sustaining the universe, for example), many contemporary disputes and approaches find their roots, or at least have parallels, in much older disputes and approaches. The thought that an account of causation should draw on our best science, for example, drives contemporary causal process theories; but it also motivated Hobbes and Boyle, and, in the first half of the twentieth century, Frank, Schlick, and Reichenbach. The thought that there is something problematic about mental causation was well known to Descartes, and the idea that causation derives from the essential natures of substances can be found in Leibniz as well as in contemporary accounts of causal powers.

Chapter 1: The Ancient Greeks

Sarah Broadie discusses the role of the Greek noun *aitia* in ancient Greek philosophy. In the theoretical inquiries of ancient Greek philosophers, to ascribe to *X* the status of *aitia* of *Y* is to imply that *Y* stands in a certain relation *R* to *X* and that this information entitles one to

claim that *Y* is understandable in the light of *X*. For Aristotle the relation *R* might be the relation of *being composed of,* of *being for the sake of,* or *being made (or brought about) by*. For Plato *Y* bears the required relation to *X* just when *X* is a Form and *Y* participates in the Form *X*. Broadie examines the role of two historically dominant models of explanation: that of demonstration from first principles and that of purposive agency. On the first model, to explain some fact is to deduce it syllogistically from principles that are necessary and self-explanatory. To explain some object on the second model is to provide information about how it is composed of certain matter arranged in accordance with a particular structure by an agent for the sake of certain end.

[Chapter 2](): The Medievals

John Marenbon discusses some issues about causation that were debated in Arabic, Jewish, and Latin philosophies over a period of a thousand years. He examines Aquinas' puzzling distinction between chains of accidental causes and chains of essential causes, as used in his Second Way. Marenbon traces this distinction back to the Persian philosopher Avicenna's doctrines about metaphysical causation, which concern the way existence is created by being passed down through a causal chain from God, who is the one necessary existent. He also examines the occasionalist traditions of thought in Islamic and Christian philosophies, especially in the work of the eleventh-century Islamic philosopher al-Ghazālī and of the fourteenth-century Parisian philosopher Nicholas of Autrecourt. This tradition has interesting links with modern treatments of causation in denying the existence of causal efficacy or causal necessity in nature. Finally, Marenbon briefly discusses the early medieval discussions of causation of Eriugena and of Abelard, as illustrations of the ingenuity of medieval thought about the subject before the rediscovery of Aristotle's work in the thirteenth century.

[Chapter 3](): The Early Moderns

Kenneth Clatterbaugh characterizes philosophers' work on causation in the early modern period as 'house cleaning': a tidying-up of the abundance of causes—material, final, formal, and efficient causes, material and immaterial causes, and God (as both the creator and sustainer of the universe)—bequeathed to them. The central dispute in this period can be seen as one between a materialist, mechanistic, and naturalist conception of causation (Hobbes, Gassendi, Boyle, Newton), according to which the behaviour of objects is to be explained solely in terms of empirically available features of matter such as size, shape, and motion; and more 'metaphysical' conceptions of causation, which appeal to God (Malebranche), the internal natures of substances (Leibniz), and so on.

[Chapter 4](): Hume

As Don Garrett notes, there is considerable disagreement amongst commentators

concerning what Hume's theory of causation really amounts to. Traditionally, Hume has been seen as espousing a version of the regularity theory of causation: causation is no more than the constant conjunction, or regular co-instantiation, of two types of events, so that *'c* caused *e'* means 'events similar to *c* are constantly conjoined with events similar to *e'*. However, Hume has also been characterized as a causal realist, for whom causation is a real relation linking events that transcends mere constant conjunction, and as a projectivist, so that the idea of necessary connection (which, according to Hume, derives from our habit of inferring that an event like *e* will occur on witnessing the occurrence of an event like *c*) plays a role in the content of causal claims while failing to refer to any mind-independent relation. Garrett presents a novel alternative view, derived from an account of Hume's conception of abstract ideas in general (of which causation is one). On this view, the impression of necessary connection is a kind of 'causal sense' by means of which we identify causal sequences, but the idea of necessary connection itself is not part of the meaning of 'cause'.

[Chapter 5](#): Kant

In his *Critique of Pure Reason,* Kant claimed that a principle of causality is necessary for any coherent experience of the world, and as such it had the status of a synthetic a priori truth. Kant's treatment of causation became the model for his treatment of metaphysics in general. Eric Watkins's chapter traces the development of Kant's thinking on causality from his pre-critical period, in which he was responding to earlier figures such as Leibniz and Wolff, through his treatment of causality in the *Critique of Pure Reason,* to his application of causality to the natural sciences in later works.

[Chapter 6](#): The Logical Empiricists

Michael Stöltzner provides a detailed account of the theories of causation of three members of the Vienna and Berlin Circles: Frank, Schlick, and Reichenbach. He notes three major philosophical influences on their views—Mach's reinterpretation of Hume, neo-Kantianism, and the conventionalism of Duhem and Poincaré—but focuses on the contemporary developments in physics (namely relativity theory and quantum theory) and probability theory, with which they were intimately acquainted. The picture that emerges of the logical empiricists' attitudes towards causation is one that has little in common with the crude verificationism with which they are often associated. On the contrary: their commitment to replacing traditional metaphysics with a scientific worldview required the development of sophisticated accounts of causation that aimed to accommodate the scientific developments of the early part of the twentieth century.

PART II: STANDARD APPROACHES TO CAUSATION

The chapters in Part II survey the leading theories of causation that have been the focal points of contemporary debate. The regularity and counterfactual theories described in the first two chapters may be said to have their origins in Hume's two definitions of causation, while the probabilistic, causal process, and agency/interventionist theorists described in the other chapters are of more recent origins. These theories take as their starting point some characteristic feature of causation—that causal relations instantiate regularities, involve continuous processes or ma-nipulability, or are marked by relations of counterfactual dependence or probabilistic relevance. The theories then attempt in their various ways to provide reductive analyses of causation in terms of these characteristic features. The chapters in this Part explore in detail the successes and failures of the theories. Which of them is the best reductive theory of causation is still a live question that is much debated.

Chapter 7: Regularity Theories

Stathis Psillos discusses the 'regularity view of causation' (RVC), according to which for one event to cause another is, roughly, for the two events to instantiate a regularity (only 'roughly' because this crude view needs to be made more sophisticated in various ways if it is to stand any chance of being a viable account of causation). He sketches the history of RVC, from the early nineteenth-century Scottish philosopher Thomas Brown, through John Stuart Mill and John Venn, to J. L. Mackie. He brings out the relationship between RVC and regularity-based accounts of laws and explanation, and defends RVC against objections concerning the notion of similarity on which it depends (since the existence of a regularity presupposes the specification of some type under which the instances of the regularity fall) and the fact that causation is an asymmetric relation. Psillos himself is a causal pluralist, but regards RVC as 'the next best thing'.

Chapter 8: Counterfactual Theories

L. A. Paul examines counterfactual theories of causation, starting with David Lewis's classic theory. According to this, causation is to be understood in terms of counterfactuals of the form 'if event c had not occurred, event e would not have occurred'. When such a counterfactual is true, Lewis says the event e causally depends on the event c; and when c and e are wholly distinct events that are linked by a chain of such causal dependences, he says c is a cause of e. After reviewing some general methodological issues concerning conceptual and ontological analysis, Paul outlines the merits of this theory by showing how it applies smoothly to examples involving common causes, early pre-emption, and causation involving absences. She examines possible solutions to especially recalcitrant problems arising from certain kinds of pre-emption and overdetermination.

[Chapter 9](): Probabilistic Theories

Probabilistic theories of causation have appealed to writers who think that causation is compatible with indeterminism. The central idea behind early probabilistic theories is that causes raise the probability of their effects. Jon Williamson's chapter traces the development of this theory in Reichenbach, Good, and Suppes. It also covers more recent work in causal modelling, in which causal relations are represented by Bayes Nets. Several difficulties with probabilistic theories are addressed. The chapter concludes with a discussion of the author's own proposal for a probabilistic theory of causation: epistemic causality.

[Chapter 10](): Causal Process Theories

Causal process theories try to characterize causation in terms of continuous processes and interactions between them, rather than in terms of relations between discrete events. Causal processes must be distinguished from various kinds of pseudoprocesses, which do not carry causal influence. Phil Dowe's chapter discusses the leading attempts to characterize causal processes and interactions in terms of mark transmission and conserved quantities. It also considers a variety of potential problems for causal process theories of causation.

[Chapter 11](): Agency and Interventionist Theories

James Woodward surveys manipulation-based accounts of causation, dividing them into agency theories that stress the connection between causation and human agency and interventionist accounts that focus on an abstract notion of intervention. He criticizes agency accounts on the grounds among others that a free human action is not necessarily the best way to manipulate a cause in an experimental test of causation. He specifies the conditions for an ideal manipulation of a cause, called an intervention, and explains several causal concepts in terms of interventions. For example, he says a variable X is causally relevant to variable Y if and only if (i) there are possible interventions that change the value of X such that (ii) under such interventions X and Y are correlated. Woodward defends this account against standard objections to manipulation-based accounts to the effect they are circular and that they are anthropocentric. He also explores the support that agency and interventionist accounts have from the empirical psychology of causal learning and judgement among both humans and non-humans.

PART III: ALTERNATIVE APPROACHES TO CAUSATION

All the chapters in [Part II]() concern theories that attempt to provide a definition or analysis

of causation. In Part III, we consider a number of further approaches to thinking about causation, all of which provide some kind of conceptual framework. None of them offer explicit definitions of causation. Nonetheless, by articulating the connections between causation and related concepts, as well as between different causal concepts, and by situating the concept of causation within broader epistemological and metaphysical frameworks, they all seek to shed light on the concept (or concepts) of causation.

Chapter 12: Causal Powers and Capacities

Stephen Mumford describes the history and the fundamental claims of the powers approach to causation. He argues that an ontology of powers or capacities can solve, or rather dissolve, the traditional problem of causation introduced by Hume. As traditionally interpreted, Hume saw the world as consisting of discrete, distinct existences and causation as a contingent, external relation between such existences. In contrast, a powers ontology accepts necessary connections in nature: instead of contingently related cause and effect, there are powers and their manifestations, which are distinct existences related by necessary connections. Whereas Hume conceptualized causal relations as asymmetric external relations involving temporal priority, Mumford argues that the powers approach allows for a conception of causal relations as symmetric internal relations involving no temporal priority.

Chapter 13: Anti-Reductionism

John Carroll argues for anti-reductionism about causation, that is, the view that causal facts cannot be analysed in terms of non-causal facts. He advances several arguments, for example from the directionality of causation and the pre-emption problem, and meets some objections to anti-reductionism: that it is uninformative, that it courts scepticism, that it is ontologically extravagant, and that the central intuitions of the anti-reductionist are 'feeble and foggy'. He ends by commenting favourably on the recent trend in the literature of addressing broader issues (concerning, for example, transitivity and causation by omission) rather than developing and refining reductive analyses.

Chapter 14: Causal Modelling

Christopher Hitchcock's chapter discusses a number of formal methods that have been developed in statistics and artificial intelligence for representing causal relationships and facilitating causal inferences. It examines the assumptions that these methods make about the relationship between causation and counterfactuals, interventions, and probability. It also describes the use of such tools to define more specific causal concepts, especially the concept of actual causation.

Chapter 15: Mechanisms

The term 'mechanism' literally refers to man-made machines, but it has become widely used to describe complex systems with interacting parts. Many areas of science, particularly the biological sciences, are concerned to a large extent with the discovery and articulation of mechanisms in this broad sense. Stuart Glennan's chapter describes philosophical accounts of mechanism, and the relationship between mechanisms and other causal notions.

Chapter 16: Causal Pluralism

The theories of causation surveyed in Part II all try to offer a single definition of causation. Proponents of pluralist approaches to causation maintain that this project is doomed to fail because there is no univocal concept of causation to be captured in this way. Some authors have proposed that there are two specific concepts of causation. A more radical position is that there is a vast array of specific causal concepts that bear only a family resemblance to one another. Peter Godfrey-Smith ends this chapter with the suggestion that causation is an 'essentially contested' concept—that disagreement about what counts as a cause is itself an essential feature of the concept.

PART IV: THE METAPHYSICS OF CAUSATION

Many issues about causation are closely intertwined with metaphysical questions about the composition and structure of the world. For example, the question of about what kinds of things enter into causal relations is partly a question of ontology. Whether causation relates objects, events, exemplifications of universals, or tropes depends on which of these entities are genuine existents. Similarly, the question of how causation is linked to laws of nature hinges on whether laws and causation are fundamental constituents of reality and whether they reduce ultimately to non-causal facts. Again, issues about the very nature of the causal relation itself—is it a natural relation that carves nature at its joints? What is the basis of its time-asymmetry?—are, at heart, metaphysical questions. The chapters in Part IV take up these metaphysical questions concerning causation.

Chapter 17: Platitudes and Counterexamples

Many theorists of the concept of causation take their theories to be responsive to commonsense platitudes about causation. Peter Menzies points out that most philosophers implicitly agree on one central platitude: namely, that the concept of causation is the concept of a natural relation, by which is meant a contingent a posteriori relation that is free of

normative considerations and that is objective and mind-independent in the sense of 'carving nature at its joints'. Menzies argues that this platitude is a philosopher's myth that is not supported by the linguistic or psychological evidence. He cites linguistic data about causal relata, the fragility of causes and effects, absences as causes and effects, and the distinction between causes and background conditions as posing difficulties for this philosophical view. He argues that these phenomena are best explained by a context-sensitive contrastive account of the concept of causation that is attuned to normative considerations.

[Chapter 18]: Causes, Laws, and Ontology

Michael Tooley provides a useful way of carving up the ontological space within which rival views of causation are positioned. First, is the view singularist? That is, does it take the causal relation between particular events to be more basic than the causal laws (if there are any) under which those events are subsumed? Second, is it a reductionist view? That is, do causal relations and causal laws reduce to non-causal facts? And third, is it a view according to which the obtaining of causal relations requires more than the existence of 'Humean' states of affairs—that is, states of affairs just involving the intrinsic properties of and external relations between facts or events, where intrinsic properties are restricted to properties whose instantiations do not entail the existence of any distinct states of affairs? Tooley himself defends a singularist, non-reductionist account.

[Chapter 19]: Causal Relata

Douglas Ehring considers the main candidates for the relata of causation: events, exemplifications of universals, tropes, facts, objects (including persons), and values of variables. He argues that consideration of general features of causation—spatiotemporal relatedness, intrinsicality, and transitivity—reveals that tropes are the best candidates, although he also argues that the standard arguments for requiring a unified account of causal relata are not persuasive.

[Chapter 20]: The Time-Asymmetry of Causation

Huw Price and Brad Weslake address the time-asymmetry of causation—the fact that causes typically precede rather than succeed their effects. They argue that a satisfactory explanation of this time-asymmetry must also explain the time-asymmetry of deliberation— the fact that we act for future but not past ends. They focus on counterfactual theories of causation that attempt to explain these asymmetries in terms of hypotheses about contingent features of the world, such as David Lewis's hypothesis of an asymmetry of overdetermination and David Albert's Past Hypothesis to the effect that the universe began in an extremely low entropy condition. They question the empirical adequacy of these hypotheses as well as their ability to explain the time-asymmetry of causation and deliberation. Price and Weslake also

consider the question 'why should we use time-asymmetric causal counterfactuals in deliberation at all?' in relation to the debate between causal and evidential decision theories over Newcomb problems. They argue that if one thinks of deliberation in evidential or non-causal terms, one can explain the time-asymmetry of deliberation in terms of our distinctive temporal orientation as agents. They extend this subjectivist explanation of the time-asymmetry of deliberation to that of causation, grounding each in the distinctive perspective we have on the world as agents.

PART V: THE EPISTEMOLOGY OF CAUSATION

How do we come by beliefs about the causal structure of the world, and when are those beliefs justified? A standard position in recent philosophy—stemming from Hume—is that causal relations are never perceived, but can only be inferred on the basis of past regularities. The first two chapters in this section cast some doubt on this claim. However, many causal relations are certainly inferred on the basis of statistical data (which is to say, on the basis of a kind of past regularity). The third chapter addresses the question of how, and in what circumstances, we should draw such inferences.

[Chapter 21](#): The Psychology of Causal Perception and Reasoning

Humans learn about the causal structure of the world and employ this knowledge in further processes of reasoning. David Danks's chapter surveys the psychological study of these aspects of our thought. One long-standing tradition focuses on causal perception: this is a relatively automatic reaction to the observation of certain sequences of events, which is not penetrated by information available from other sources. Another tradition focuses on the inference of causal relations on the basis of perceived regularities. A more recent idea is that we deploy a certain kind of causal model to represent causal relations. This chapter also compares human causal understanding with what is known about causal reasoning in other animals.

[Chapter 22](#): Causation and Observation

Helen Beebee discusses the role that the debate about whether causation can be observed has played in the debate about the metaphysics of causation. She argues that from a contemporary perspective, Hume's argument that causation cannot be observed is unpersuasive, and considers some rival criteria that have been offered by psychologists and philosophers for deciding whether causation can be observed, concluding that it looks as though it can be (and indeed often is). However, she argues that this conclusion has no

interesting consequences for the metaphysics of causation.

[Chapter 23](): Causation and Statistical Inference

Statistics is concerned with many different kinds of problem: the compact representation of quantitative facts, estimating parameters from partial data, testing probabilistic hypotheses, and constructing probabilistic models. Clark Glymour's chapter concerns the problem of when statistical quantities can be given a causal interpretation, and when causal relationships can be reliably inferred using the various statistical methods. It covers traditional statistical tools such as regression analysis, as well as more recent developments, such as the use of graphical causal models. It concludes with a survey of a number of puzzles about the relationship between causation and statistics.

PART VI: CAUSATION IN PHILOSOPHICAL THEORIES

A conspicuous feature of much twentieth-century analytic philosophy is the central role it accords to the concept of causation. On the one hand, philosophers have appealed to causal concepts in order to analyse or clarify a broad range of concepts such as knowledge, perception, free action, and semantic content. On the other hand, the concept of causation has also figured centrally in many conceptual puzzles explored by philosophers. For example, ethicists have considered whether there is a distinction between killing and letting die, and whether an act is morally permissible even if it has foreseen bad effects. Philosophers of mind have asked whether mental states can have causal effects independently of the brain states on which they supervene. And philosophers of science have asked whether all explanation is causal explanation and whether higher-level theories that fail to reduce to lower-level theories allow for emergent causal structures. All these questions turn on subtle issues about the nature of the causal concept and its relationship to other philosophically significant concepts. The chapters in [Part VI]() take up these issues in detail.

[Chapter 24](): Mental Causation

Cei Maslen, Terry Horgan, and Helen Daly survey six philosophical arguments that purport to demonstrate that mental causation does not exist; that is, that mental states and properties are not causally effective. Many of these arguments turn on the thought that since the physical properties on which mental properties supervene do all the causal work, mental properties are causally irrelevant. Maslen, Horgan, and Daly argue that traditional responses to these arguments that posit type identities between mental and physical properties are not successful. They advance a new approach to these problems in terms of a contextualist theory

of causation, according to which the truth or falsity of causal claims depends on which possible worlds are deemed relevant in a given context of inquiry. They explain how this contextualist theory answers each of the six arguments that threaten mental causation.

[Chapter 25](#): Causation, Action, and Free Will

Alfred Mele discusses issues concerning causation as they appear in the debates about action and free will. In the case of action, the central debate has been between causal and non-causal theories of intentional action, with causalists claiming that intentional actions are characterized by their being caused (in the right way) by the agent's reasons (or belief/desire pairs), and anti-causalists claiming that while actions are done *for* reasons, they are not caused *by* reasons. Mele defends causalism against two standard problems: causal deviance and vanishing agents. In the case of free will, a version of the classic problem of free will is the worry that if agents' choices are deterministically caused by their beliefs and desires, then they cannot choose differently, and so lack free will. On the other hand, Mele urges that the view that agents' choices are indeterministically caused by their prior mental states suffers from a 'problem of luck': the indeterministic causation seems to prevent the agent from having any genuine control over which choice she makes. Mele also argues that agent-causalism, the view that freely made decisions are caused by the agent, and not by her mental states, is also subject to a problem of luck.

[Chapter 26](#): Causation and Ethics

Carolina Sartorio investigates the role of causation in various theories and issues in ethics —in particular consequentialism, the distinction between killing and letting die, the doctrine of double effect (according to which an act can be morally permissible even if it has foreseen bad consequences), and the concept of moral responsibility. She considers the extent to which various problems in ethics can be resolved by paying attention to the concept of causation. For example, a standard objection to consequentialism is that an act will in general have far too many distant consequences for them all to be relevant to its moral status; Sartorio argues that denying the transitivity of causation provides a way out of this problem. By contrast, she argues that the moral difference between killing and letting die probably cannot be accounted for in terms of a difference in causal status between acts and omissions.

[Chapter 27](#): Causal Theories of Knowledge and Perception

Ram Neta observes that philosophers have appealed to causation in order to explain what it is for someone to *perceive* an external object and what it is for someone to *know* about external things by means of perception. (Confusingly, the expression 'causal theory of perception' has been applied to both these kinds of explanations.) Both explanations proceed in terms of a causal relation existing between an external entity and a mental entity. Neta

enumerates the different kinds of things that theorists have identified as the external and mental entities in the two cases. He concludes by considering a number of criticisms that have been lodged against causal theories of empirical knowledge.

[Chapter 28](): Causation and Semantic Content

Frank Jackson reviews the way in which causal concepts have figured in accounts of the content and reference of thought and language. He critically examines informational treatments of content, according to which the content of a thought is the state of the world that co-varies with the thought; and selectional treatments according to which the content is the state of the world that is selected to co-vary with the thought. He notes problems for both approaches. He then considers the content of language, focusing on the reference of proper names. He outlines the description theory of names, examines Kripke's influential criticisms of this theory, and defends the theory against these criticisms as well as a separate objection concerning reference in symmetrical worlds. He shows how a description theory can incorporate the main elements of Kripke's alternative causal-historical theory, according to which reference is secured through information-preserving chains of reference-borrowing.

[Chapter 29](): Causation and Explanation

Causation and explanation seem intimately connected. We often explain why some event occurred by showing how it was caused. Peter Lipton's chapter explores the connection between causation and explanation. It asks whether all explanations cite causes, and whether all causes are explanatory. It also examines why causes explain their effects, but not vice versa. It concludes with a discussion of how explanatory considerations facilitate causal inference.

[Chapter 30](): Causation and Reduction

Paul Humphreys addresses two issues in this chapter: the issue about how causation operates between systems that stand in reductive relations to one another; and the issue of whether causation is itself amenable to a reductive treatment. He says that the basic idea behind reduction is that if X can be reduced to Y, X is nothing but Y. More specifically, on Ernest Nagel's model of theory reduction, one theory reduces to another theory when the laws of the first theory can be deduced from the laws of the second with the aid of bridge laws. Some contemporary philosophers question this model, in particular the bridge laws it posits, on the grounds that the higher-level properties are often multiply realized by lower-level properties. Humphreys challenges some claims of multiple realizability, arguing that they must be assessed on a case-by-case basis. He examines two arguments—the exclusion argument and the downwards causation argument—that purport to show that there is no causation outside the physical domain; and considers possible emergentist responses to these

arguments. Finally, he considers an alternative model of reduction developed by Jaegwon Kim to address the issue of emergence.

PART VII: CAUSATION IN OTHER DISCIPLINES

Causation enters many fields of inquiry in a wide variety of ways. Some fields are centrally concerned with discovering causal relations. This is particularly true of applied fields such as epidemiology, agronomy, and forensic pathology, but even more theoretical areas of biology and physics concern themselves with the discovery of causal relations. In physics, these causal relationships are often dubbed 'effects': the Hall effect, the Lamb-shift effect, the Zeeman effect, and so on. Many fields also deploy specific concepts that are explicitly or implicitly causal. In the law, one talks about 'proximate' causes, while geneticists frequently search for 'specific' causes. In evolutionary biology, a number of authors have argued that concepts such as 'fitness', 'function', and 'selection for' are to be understood in causal terms. In addition, some theories in fundamental physics seem to have consequences for the very nature of causation, or for what kinds of causal relations can be found in the world.

[Chapter 31](): Causation in Classical Mechanics

Classical mechanics comprises the research programme in physics pioneered by Newton and continuing through the work of Maxwell. Although it has been supplanted by the theories of relativity and quantum mechanics, classical mechanics is still widely used as a successful approximation. The laws of classical mechanics take the form of differential equations, and there has been considerable debate as to whether these laws lend themselves to causal interpretation. Russell famously argued that causes played no role in (classical) physics. Another issue addressed in Marc Lange's contribution concerns the role of locality principles, especially in electromagnetic theory. While the laws of classical mechanics are not explicitly probabilistic, a number of authors have claimed that classical physics admits cases of indeterminism. A final topic discussed in this chapter is the role of least action principles in physics, and whether these involve a form of teleology.

[Chapter 32](): Causation in Statistical Mechanics

A number of authors have attempted to ground the asymmetry of causation in the temporal asymmetry captured by the second law of thermodynamics, but there are other important issues about the relationship between causation and statistical mechanics. Larry Sklar's chapter explores explanations in statistical mechanics that appear to be non-causal. The first is

a kind of transcendental explanation that is used in phenomenological equilibrium thermodynamics. The second involves explanations of the approach to equilibrium in non-equilibrium thermodynamics. These explanations make use of a probability distribution that appears to be wholly a priori.

[Chapter 33](): Causation in Quantum Mechanics

Quantum mechanics seems to tell us that the behaviour of the particles at the microscopic level violates our expectations for the behaviour of physical bodies derived from our experience with the macroscopic world. In particular, quantum mechanics seems to require the failure of causal determinism, and seems to admit non-local causal influences. Yet despite its empirical success, there is still widespread disagreement about how to interpret quantum mechanics, that is, about how to understand what quantum mechanics is telling us about the behaviour of the world at the most fundamental level. Richard Healey's chapter explores the consequences of these debates for our understanding of causation.

[Chapter 34](): Causation in Spacetime Theories

Newton's physics, Einstein's special theory of relativity, and Einstein's general theory of relativity all have profound implications for the structure of space and time. The structure of space and time, in turn, constrains what kind of causal relationships can hold. Carl Hoefer's chapter explores these connections. Newton's conception of absolute space and time admitted an absolute relation of simultaneity. One consequence of this structure is that the theory imposed no upper bound on the velocity of causal processes. Indeed within the framework of Newton's theory, gravitational attraction seems to involve instantaneous causation at a spatial distance. Einstein's special theory of relativity, by contrast, seems to impose an upper limit on the velocity of causal influences. This upper limit is captured in the light-cone structure of Minkowski spacetime. There is thus an intimate connection between the structure of spacetime and the possibility of causal influence. Einstein's general theory of relativity admits many spacetime structures with apparently strange causal behaviour. Some contain closed causal curves, and thus allow for a kind of time-travel or self-causation. Others contain regions that are forever causally isolated from one another. Moreover, the general theory of relativity might be read as indicating that spacetime itself causally interacts with matter.

[Chapter 35](): Causation in Biology

Samir Okasha examines the role of causal concepts in biology, especially evolutionary theory and genetics. After providing a historical survey of the contributions made by biologists to the understanding of causality, he looks at causation in evolutionary theory. In clarifying the causal structure of evolutionary explanations, he draws on some distinctions introduced by Sober between developmental and selectional explanations and between

selection for and selection of traits. He also employs Ernst Mayr's distinction between functional and evolutionary biology to elucidate the sense in which evolutionary explanations are teleological. He concludes by looking at the causation in genetics. He clarifies the sense of the expression 'the gene for' in modern genetics and examines the controversy over heritability analyses of cognitive traits such as IQ and the critique of the centrality of genetic explanations by developmental systems theory.

[Chapter 36](): Causation in the Social Sciences

Harold Kincaid discusses a number of general issues concerning causal claims in the social sciences. One issue relates to the criticism that causal claims in the social sciences are untestable because they must be qualified by *ceteris paribus* clauses, and the criticism that they do not apply to social phenomena because of their aggregate or non-individual character. Kincaid argues that these criticisms are misconceived. Another issue relates to the various distinctive kinds of causes and explanations that figure in the social sciences, especially teleological causes and functional explanations. Kincaid provides an account of functional explanation as a special kind of causal explanation that does not make any illicit appeal to biological analogies. The final issue he discusses relates to whether the assumptions made by causal modelling techniques fit with much social science research. He defends the validity of case studies and case comparisons as ways of responding to the fact that causal modelling assumptions are inapplicable to many problems in social science.

[Chapter 37](): Causation in the Law

Jane Stapleton provides a wide-ranging critique of existing accounts of causation in the law, including those of the early US realists, Hart and Honoré, Michael S. Moore, the lawyer-economists, and Richard Wright. Her central contention is that the notion of 'cause', as used in legal contexts, must be 'untainted by normative controversies'. This is something that many existing accounts fail to achieve, for example by using the notion of a 'proximate cause', as though there were some factual, worldly distinction between causes that are and are not proximate rather than a normative decision concerning the extent of legal liability. Appeals to commonsense intuitions about the meaning of 'cause' are also, she argues, liable to smuggle in normative features. Stapleton sketches a new account, tailor-made to fit the uses to which causal language is legitimately put in legal contexts. The aim is to achieve an understanding of causation such that 'where all the facts of the case are known' there is 'no room for disagreement on the issue of "causation"'.

PART I
THE HISTORY OF CAUSATION

CHAPTER 1
THE ANCIENT GREEKS

SARAH BROADIE

Felix qui potuit rerum cognoscere causas.

(Blessed accomplishment theirs, who can track the causes of things.)

Vergil, *Georgics* 2. 490

1. A Multidimensional Concept

We must begin by looking at the Greek noun, *aitia* and its cognate adjective *aitios, aition.*[1] *Aitia* is traditionally translated 'cause', although many prefer 'reason' or 'explanation'. The words originally ascribed responsibility with a view to faultfinding or giving credit. In this early usage, the noun refers to a certain property or status, and the adjective picks out status-holders. The status is commonly that of being guilty of, or being charged with or blamed for, something bad; occasionally it is that of deserving thanks for something welcome. So: to call a person or object *X aitios* or *aition* of an event, action, or state of affairs *Y* is to assign to *X* the above status (deplorable or laudable) in relation to *Y*. It is not simply to claim that *X* stands to *Y* in the relation of having done or brought it about, although this is implied. It is also to signal that the community's response to *Y* should involve the adoption of certain attitudes and actions towards *X*. In short, it is part of the meaning of the noun and the cognate adjective that they are assigned or ascribed to someone or something, *X*, in response to the presence, existence, or happening of something, *Y*, from the perspective of a certain kind of interest: crudely, interest in *whom or what to blame or thank for Y*.

These terms came to figure in the more impersonal, less judgemental, more theoretical inquiries of historians and philosophers. They continued to presuppose a perspective of interest from which they were applied, but it was now an interest in *systematic explanation.* To assign or ascribe to *X* the status of *aitia* of *Y* is now to signal to those interested in getting to understand *Y*, or make sense of it in a scientific and systematic way, that they should look upon it as related somehow to *X*. What is this relation *R* in which they should regard *Y* as standing to *X*? We may be tempted to gloss it as 'caused by'; but if 'cause' is our translation of *aitia* and cognates, the result will be circular. How then to specify *R?* Here are some examples. For Aristotle, for some substituends of '', *R* is the relation: *- is composed of-*. For some, it is: *- is for the sake of-*. For some, it is: *- is made* (or: *brought about) by -*. We shall look more fully at the Aristotelian picture soon, but meanwhile consider also his teacher

Plato's notorious method of explaining (assigning the *aitia* of) this woman's beauty, or my bicycle's wheels' duality: the transcendent Form of the Beautiful, or of the Two, is given the '*X*'-position, and the woman or my wheels are said to stand to it in the relation – *participates in* –.[2]

So the ascription to *X* of aitia-status in relation to *Y* has implications on two levels. It implies (1) that *Y* stands in some specified relation, *R*, to *X*; and (2) that being in possession of this information gives or entitles one to claim, concerning *Y*, a certain attitude or cognitive position that points at *X*. This is so whether the context is the original juridical one, or the theoretical, explanatory, one.[3] Hence an ascription of aitia-status can be doubly challenged: by questioning the basic factual claim ('Did *X* do or bring about *Y*?', 'Is *X* as distinct from some other material what *Y* is composed of?'); and, granting the basic factual claim, by asking whether its truth has the significance that matters in the context. Thus the juridical perspective recognizes circumstances where even if *X* did bring about/do *Y*, it is inappropriate to award *X* blame or credit for that. Similarly, ancient Greek philosopher-scientists could agree that *Y* is composed of *X*, yet violently differ on whether this is what makes sense of, is to be held up as the systematic explanation of, *Y*. The explanatory context can also create the quite different situation where rational belief that X-like objects even exist is in thrall to their supposed power to explain Y-like objects. In this way Aristotle famously rejected the reality of Plato's Forms, along with the metaphysical participation-relation in which ordinary things are supposed to stand to them.

Thus it was not accidental, but the result of the multidimensional meaning of the word *aitia*, that ancient Greek debates over 'the causes of things' were not only about the identity of the causes, but also, and more fundamentally, about what counts as explanation. In fact, Aristotle's critical surveys of earlier theorists show these pioneers on the whole failing to distinguish the two levels of inquiry: the empirical or scientific level, as we would call it, and the philosophical and conceptual one where the questions belong to metaphysics and epistemology. Plato is our first great witness on how intertwined these levels can be. He shows this by disentangling them himself, perhaps for the first time ever and unaided by any pre-labelled contrasts such as 'empirical' and 'conceptual', 'scientific' and 'philosophical'.

This unfolds in the portion of Plato's *Phaedo* known as 'Socrates' intellectual autobiography' *(Phaedo* 96a-100b, especially 97b-99c). Conversing with friends on his day of execution, Socrates tells how as a young man he was fascinated by questions of biology, physiology, and psychology, but bewildered over how to answer them. One day he heard a public reading from the treatise of Anaxagoras, the great fifth-century philosopher-scientist. It was taken from where Anaxagoras declares cosmic *intelligence* to be what causes and orders everything. Socrates immediately expected the treatise to proceed not only to answer such standard questions as whether the earth is convex, and whether it lies at the centre, but also to expound the 'cause and necessity' whereby *it is better thus and so*. For intelligence seeks what it values (and cosmic intelligence, presumably, values only what is truly better). So: to cast *intelligence* for the role of 'fundamental force of nature'—a move that verbally resembles casting *gravity* for it, or *heat* or *electromagnetism*—is not simply to claim that the cause is so and so rather than such and such: it is also to commit oneself to an entire style of scientific explanation—the teleological style.

Anaxagoras (if we accept this impression of him in the *Phaedo)* cannot have been clearly

aware that he was shouldering this commitment, for his detailed explanations failed to honour it. Socrates on perusing the work was bitterly disappointed to find the body of it appealing only to mechanical and materialistic factors such as air and water and the vortex. These, says Socrates, are mere necessary conditions: they are not 'real causes' at all, because citing them alone fails to show why it is good that the actual phenomena exist rather than others. In short, for Socrates in the *Phaedo* and for Plato its author, if we substitute 'intelligence' for '*X*' in '*X* is the *aitia* of *Y*', the result is not yet a *scientific explanation* of *Y,* a specific explanation such as a scientific expert would offer; instead, it implicitly announces the form a genuine scientific explanation of *Y* should take: it should show that *Y* exists or happens because *Y* is an intrinsically good or beautiful object or state of affairs, or because it contributes to such. Anaxagoras' failure to spell out such explanations for specific phenomena was, according to Socrates, a complete failure to 'use his intelligence'—both his own and that which his cosmology theoretically postulated *(Phaedo* 98b8–9).

Anaxagoras, of course, was aiming to be scientific, but scientific explanations work only on the basis of coherent and shared presuppositions concerning the correct general form of scientific explanation. To introduce intelligence as universal cause and then leave it hanging idle in the theory betrays lack of clarity at this fundamental level. Anaxagoras as Socrates presents him had scientific ambitions but no clear assumptions as to the sort of theory that would satisfy them, nor any awareness that the laying out and justifying of such assumptions is a huge task in itself.

One thing we learn from 'Socrates' autobiography', then, is that when we see ancient Greek philosopher-scientists plunging themselves into questions about *the causes of things*, we have to look at each case to decide whether they are seeking specific explanations according to some less or more conscious and clear-cut general model of explanation; whether they are enquiring about what the general model should be; or whether they are operating indiscriminately on both these levels, and therefore not very effectively on either. Below I shall describe two explanatory models that have dominated historically: that of demonstration from first principles, and that of purposive agency. The main reference point will be Aristotle. Both conceptions were hatched earlier (by axiomatizing mathematicians in the one case, by Plato above all in the other), but Aristotle gave them their most serious philosophical development.

2. EXPLANATION BY DEMONSTRATION

One can prove deductively from true premisses *that* something is the case. Take an example that Aristotle made famous: if one knows that (1) <u>astronomical objects are divided into those that are near (to the earth) and do not twinkle, and those that are not near and do twinkle</u>, and (2) <u>the planets do not twinkle</u>, one can obtain the new information that (3) <u>the planets are near</u> (by comparison with the fixed stars). In this argument (A), we start with a question of fact: 'Are the planets near?', and discover by deduction that an affirmative answer is correct. But we have not explained or shown the cause of anything. On the other hand, suppose that the initial question—given that (2) <u>the planets do not twinkle</u>—is 'Why is this so?' We can now construct an argument (B) using the relevant part of premiss (1) together with the newly acquired information (3) to conclude that (2) <u>the planets do not twinkle</u>; and in

so doing we have explained this fact, given its 'Why?', given its *aitia* or cause. Explaining why something is so simply is deducing it, or showing another how it follows, from suitable premises. In moving in this way from (1*) [the relevant part of (1)] <u>what is near does not twinkle</u> together with (3) <u>the planets are near</u> to (2) <u>the planets do not twinkle</u>, we have explained why the planets do not twinkle: it is because they are near. This explanation is an example of what Aristotle means by 'demonstration' *(apodeixis)*, and he defines scientific knowledge *(epis-tēmē)* of *p* as the ability to demonstrate *p*.[4]

Notice that the explanans in B, conveyed by the term <u>near,</u> is absent from the conclusion (which is as it should be, since the conclusion is the explanandum) and present in each of the two premises. In the terminology of syllogistic logic, of which the valid arguments A and B are informal illustrations, <u>near</u> in B is the *middle* term, and <u>planets</u> and <u>non-twinkling</u> the *extremes*. Every valid syllogism links the extremes *via* the middle term, this linkage being expressed in the conclusion, where one extreme is predicated of the other. For example, the middle term of A is <u>non-twinkling</u>,[5] and the extremes are <u>planets</u> and <u>near</u>. But here the truth of the conclusion is not being *explained* by the conjoint truth of the premises. This is so for two reasons. First, explanation can be given and received only for what, when the premises are assembled, is already assumed to be a fact or a truth not standing in need of justification: but in A the premises have been put together in order to find out whether the planets are near, or to justify to others the claim that they are. Secondly, it strikes us as obvious that the non-twinkling of the planets cannot be what explains their proximity to the earth: that the proximity is not an effect of the non-twinkling but more likely to be its cause. So, in A, the way in which the middle, <u>non-twinkling</u>, occurs in each of the premisses enables *us to discover* the linkage of the extremes that constitutes the conclusion, but it is not as if this middle represents that real property, feature, or 'nature' which, independently of our cognitive successes and failures, unites the real properties or natures represented by the extremes. By contrast, this is exactly what we get with argument B. Here, the middle, <u>near,</u> stands for that real feature of the planets whereby they are non-twinkling. This feature can be thought of as 'underlying' or 'underpinning' the non-twinkling. Such metaphors suggest both that the nearness is what is responsible, in planets, for non-twinkling, and that it is more fundamental: the latter by comparison is a 'surface feature'.

Explanatory demonstration is a matter of formal logic in that it employs the logic of the syllogism. But logic cannot reveal premisses from which to construct arguments useful for such demonstration or for any other purpose. Still, we can list characteristics that explanatory premisses should possess. They should, Aristotle says, be true and necessary, since scientific knowledge is of what cannot be otherwise; and what cannot be otherwise, if dependent at all on a cause, must depend on what cannot be otherwise.[6] The premisses must also be 'prior to', 'more intelligible than', and 'causative of' the conclusion *(Posterior Analytics* 1. 2). The priority is cognitive: *p* is prior to *q* if a correct grasp of *p* requires a correct grasp of *q*, but not conversely. *p* is more intelligible than *q* if *q* stands in need of explanation more than *p* does. For Aristotle, these cognitive rankings and dependencies are themselves to be explained by the causal dependence of what the conclusion (considered as a statement) states on what the premisses state. Ideally, explanatory demonstration starts from premisses that are 'primary' and 'immediate': that is, the linkage of their terms depends on no middle term. Thus the ideal

premisses are themselves indemonstrable. They are, or stand for, uncaused causes, hence are primary in the system of explanation. Just as it strikes us as obvious that argument B gives an explanation whereas A does not, so it probably strikes us as obvious that the premisses of B present linkages that themselves demand to be explained by a more recessive middle term. B, then, takes us only one step back in the direction of the fundamental. A full explanation of the planets' twinkling would consist in several converging series of demonstrative syllogisms starting, each, from indemonstrable premisses and showing along the way why the planets (literally the 'wanderers') are nearer Earth than the fixed stars, and also why (for celestial objects) twinkling is bound up with remoteness.

The demand for indemonstrable starting points is for primitive linkages that not merely happen to lack middle terms, but essentially pre-empt any such connector by being *self-explanatory*. That is: a mind fully attuned to such a fact would be completely unmoved to wonder why it is so. Its attitude would necessarily be: 'Of course'. Curiosity or puzzlement is as wrong-headed here as a boorish person's *lack* of curiosity about causes of the caused facts lower down the explanatory system. However, it is one thing to expound in general terms this notion of an explanatory starting point, quite another to give plausible illustrations and justify particular claims to the status. Aristotle has been read as saying that we possess a mysterious faculty of immediate, infallible, intuition that apprehends the starting points. Fortunately, his remarks do not demand this philosophically unattractive interpretation; but the ingredients of a positive alternative account are obscure and controversial.

Such difficulties prompt the question, anyway for empirical phenomena: why assume that explanation must be rooted in uncaused causes? Aristotle argues that it cannot go round in a circle. We can agree on that, since explanation, or least the sort with which Aristotle is concerned, involves the asymmetric, transitive, relation -*is more fundamental than* -: and nothing can be more fundamental than itself. He also argues that it cannot go back ad infinitum, since a demonstration can only be accomplished through a finite number of steps *(Posterior Analytics* 1. 3). We can agree again but still ask why there may not be demonstrative explanations that genuinely illuminate their explananda even though starting from premisses that themselves demand explanation. One can also query the assumption that deriving a datum D from indemonstrable starting points via intermediate premisses necessarily explains D better than if we had simply derived it from the intermediate premisses. Suppose that as well as having the explanation (E_1) of the planets' non-twinkling in terms of their nearness, we also had the explanation (E_2) of their nearness in terms of deeper starting points that need no explanation: does that make E_1 a better explanation than if we possessed it without E_2? Aristotle would certainly say yes: E_2 completes E_1, giving the planets' non-twinkling its full causal genealogy.[7] This assumption that things have, and that science should seek out, *primary* causes is also a feature of our second explanatory model, to which we now turn.

3. THE PURPOSIVE-AGENCY MODEL

We must proceed to consider causes, their character and number. Knowledge is the object of our inquiry, and we do not think we know a thing till we have grasped the 'why' of it (which

is to grasp its primary cause). So clearly we too must do this as regards both coming-to-be and passing away and every kind of natural change, in order that, knowing their principles, we may try to refer to these principles each of our problems. (1) In one way, then, that out of which a thing comes to be and which persists, is called a cause, e.g. the bronze of the statue, the silver of the bowl... (2) In another way, the form or the archetype, i.e. the formula of the essence [of the thing being explained], is called a cause (e.g. of the octave the relation of 2:1[8]...). (3) Again, the primary source of the change or the coming to rest; e.g. the man who deliberated is a cause, the father is cause of the child ... (4) Again, in the sense of end *(telos)* or that for the sake of which, e.g. health is the cause of walking about. ('Why is he walking about?' We say: 'To be healthy', and having said that we think we have assigned the cause.) The same is true also of all the intermediate steps which are brought about through the action of something else as means towards an end, e.g. reduction of flesh, purging, drugs, or surgical instruments are means towards health. All these things are for the sake of the end, though they differ from one another in that some are activities, others instruments. (Aristotle, *Physics* 2. 3, 194b16–195a3, *Revised Oxford Translation* with minor changes)[9]

Metallurgy illustrates the way in which, for Aristotle, all the types of cause fit together, so that someone who wants a full explanation of *X* must identify all of them in *X*'s case. With metallurgy, *X,* the thing to be explained, is the statue or bowl, or the coming-to-be of the statue or bowl. (These formulations are virtually synonymous, since artefacts and anything analogous have essentially come-to-be.) So the complete account of *X*, say a bowl considered as an object that has come-to-be, places it in a pentadic relation as follows: *X* is composed of matter *M* arranged in accordance with structure *S* by agent *A* for the sake of end *E*. Thus, for example, a particular bowl is fully understood when it is seen as made of silver arranged as a hollow flattened hemisphere by a metalworker for the sake of serving food at festivals. This full understanding must not only be comprehensive, specifying all the types of cause that are present (i.e. assigning specific values for the variables '*M*', '*S*', '*A*', and '*E*' for a given value of '*X*'), but must also accurately distinguish the types, recognizing that they are not in competition with each other, that none is redundant, and that they are interdependent. For example, *X* cannot have a that-for-the-sake-of-which, *E*, unless there is an agent, *A* (actual or potential), to produce *X* for the sake of *E*; and there cannot be an *A* of *X* unless there is an *E* for the sake of which *A* produces *X*. Similarly, silver cannot function as matter, *M,* of *X,* or the hollow hemisphere as the structure, S, of *X,* unless there is or has been an *A* to impose that *S* on that M. If from this Aristotelian perspective we look back at Socrates' complaint against Anaxagoras, we see that Anaxagoras fell short by specifying an agent, namely Intelligence, for the various arrangements of the cosmos, but failing to specify the end for-the-sake-of-which each arrangement was instituted.

But is this criticism of Anaxagoras much too demanding? Many modern scientists and naturalists have believed nature to be the product of divine intelligence and that natural arrangements are therefore good, but have modestly claimed for science the sole task of investigating the arrangements themselves and their workings, leaving questions of what they are good *for* to religion or theology to raise, let alone answer. Aristotle could reasonably reply

that this attitude simply assumes the view that he rejects: that we can have a full scientific explanation of something that (we believe) has a purpose without knowing what that purpose is. He could support his rejection by many illustrations, beginning perhaps with one that Plato ascribes to Socrates in the *Phaedo:* we simply have not explained Socrates' behaviour—the fact that he sits in prison awaiting execution (which he chose to do, since his escape could easily have been arranged)—if we speak only of the mechanics of sitting (angles and connections of bones, tendons, and muscles) and omit to state his purpose: to carry out the sentence of the Athenian state *(Phaedo* 98c2-e5). We simply have not explained the heart, or the lungs (it could even be said that we do not know what these organs are) if we supply no picture of the biological purposes they serve. In fact, the purpose in each case helps shed light on the other causal factors, providing clues for identifying the structures and materials of these organs, and explaining why these rather than other structures and materials are the actual ones.

Most of the illustrations in the quotation above from Aristotle are like the example of Socrates sitting in prison: drawn from the realm of rational human activity governed by thought and desire. But, as the passage indicates (see its third sentence), Aristotle's purposive-agency model is intended to apply to natural phenomena, and this is its standard application throughout his science and philosophy. Equally standard, though, is his choice of illustrations from human expertises such as medicine, building, and sculpture. The chief questions raised by this whole approach are first: if the model is for explanations of non-human natural phenomena, why does Aristotle expound it through examples drawn from human action? Secondly: what justifies his assumption that such a model, involving as it does regular recourse to purposive or teleological explanation, is in fact the gateway to understanding the natural world?

The answer to the first question is twofold. In the first place, although Aristotle is in no doubt himself that natural phenomena are governed by purpose, this position was controversial. By contrast, the claim is uncontroversial as applied to human activity. Aristotle hopes to show that it also holds of the natural world by getting his opponents to accept various analogies between natural and human activity.[10] Secondly, the distinctions between the four types of cause are most easily learned by considering artificial production. This is because in the natural cases different causal aspects often coincide, whereas in artificial production they are concretely different. It is not easy to distinguish the matter from the structure of a living organism, whereas in the case of the silver bowl, the unshaped matter was very clearly present before any useful or aesthetically interesting structure had been imposed. Again, in the case of the bowl, the agent of its shaping is external to its matter: it is a human expert, or (according to Aristotle's more philosophical analysis) it is the expert's skill considered as an active source of change. By contrast, when a living thing is shaped, say through embryological development, the force that steers this process is inseparable from the matter of the organism. Finally, the bowl is clearly an object distinct from the activity or use which is the end for which it was made: the bowl exists even when no one uses it at all. But a plant or animal is not thus clearly distinct from the active biological functioning for the sake of which, according to Aristotle, the matter of which it is made was structured into a whole of limbs, organs, living tissues, and so on. If you subtract all the biological activity, i.e. all the creature's active *use* of its organs and tissues for the purpose of staying alive, you have destroyed the organism even

though the remains may keep their shape for a while.

   The second question was: what justifies Aristotle in applying his fourfold causal scheme, which is so tellingly illustrated by human rational activities governed by thought and desire, to the world of nature? Is he simply animistically personifying natural substances and natural processes? Not at all. Aristotle's science draws boundaries between animate and inanimate; among animates, between sensate (animals) and insensate (plants); and among sensates between those that can reason and think and those that can only sense, feel, and react. Aristotle applies the fourfold causal scheme not to absolutely everything in the natural world, but to all organic systems, and he is clear that its applicability to them does not presuppose or imply mental capacities of any sort on their part. Nor does he think that its applicability rests on the assumption that the systems were designed by a divine mind. On the contrary, Aristotle's view is that each living individual is a vitalistic principle (in his terms, a *phusis* or 'nature', or a *psuchē,* often translated 'soul') expressing itself in and through a material body whose shape and organization, as well as its growth and maintenance, are effected by that principle itself, which operates as an *inbuilt non-mental skill* for realizing just that organism and its lifestyle, including its propagation. From one point of view the organism's physical parts and indeed entire physique are the *products* of its enmattered inner principle, and from another they are its *instruments* for the project of carrying on this individual's characteristic lifestyle. That is the end for the sake of which the creature's physiological parts and processes exist. Thus a full causal explanation of an organ, or of a process such as respiration, or of an anatomical arrangement, will account for its ordering and its properties, material and structural, by showing in detail what biological good the item contributes. The biological good in question is that of the very same individual whose liver, or respiratory system, or arrangement of the vertebrae, is being considered. The immediate and fundamental explanation of the presence of these items is that each contributes to the well functioning of the individual to which it belongs. Of course the individual inherited these characteristics from its forebears, but this is not the fundamental explanation of why it has them. In the Aristotelian view, things are the other way round: it is because these characteristics are to the individual's own good (under natural conditions) that it was born with them.[11]

   This teleological approach is suited to the biological domain, where thinking about objects in terms of their possible purposes in an overall system is indispensable for generating fruitful lines of research into the detail of materials and workings, and where observed features that initially seem puzzling standardly cease to puzzle once their function has been identified. But one might well ask why the methodological needs of this specialized domain should dictate the form of causal explanation in general. For that is what is going on. The purposive-agent model claims, in effect, to constitute a standard for causal explanation as such, or at least for the central, most indubitable, least watered-down type of causal explanation. In Aristotelianism this model enjoys the cachet enjoyed in later times (and in some quarters even today) by the 'mechanical' model of clashing snooker balls. Each in its epoch has been held up as the paradigm of causal intelligibility, and (in Aristotelianism at least) cases that fail to conform are generally conceptualized as derivative from the model in one way or another. But what can justify assigning this centrality to an approach that is properly at home with only a restricted range of phenomena, the phenomena of biology? Aristotle was, of course, a working scientist as well as a philosopher, and his own greatest scientific contribution was to biology.

Is it possible that this particular interest of his skewed his intellectual vision so that he saw a certain restricted domain, which especially happened to captivate *him,* as typical of the natural world in general?

At work behind these questions is the modern Western scientific picture of the world. According to it, the realm of inanimate matter is vastly more extensive in space and time than the realm of life. Even on our planet life crops up only here and there, and it has appeared here only after aeons of dramatic changes on the inanimate level. Life, moreover, depends on, and many think can in the end be completely explained by, inanimate processes. Arguably, there is no purchase for teleological explanation of specific inanimate formations and events; and yet of course they are subjects for causal explanation. Causal explanation, then, as we see it, should be tailored, typically and centrally, to cover the inanimate.

Aristotle's approach makes sense, however, in relation to his own, quite different, worldview. I shall summarize its relevant points, with the aim of showing that his predilection for the purposive-agency model of causal explanation is based on something more principled than a bias towards a limited field in which, as a matter of personal fact, he was particularly interested. Aristotle's underlying positions, although archaic to us, are reflectively reached and carefully defended.

His Earth is at rest at the centre of a spherical universe of finite diameter. All the species of living beings have always existed and will always exist. The astronomical bodies, which circulate eternally round the Earth, are living immortals. (In one place Aristotle suggests that their movement is due to a kind of longing for complete perfection.[12]) While terrestrial beings all consist of the inanimate elements, earth, water, air, and fire, the *living* individual cannot be fully explained by reference to combinations and arrangements of such materials: its composition also depends on a distinct organizing, animating, principle of form.[13] Metaphysically speaking, living organisms are more truly substances, more complete examples of what it is to be real, than inanimate objects are. This is because the former are actively responsible for their own existence and their individuality by the way they organize their own matter and carve out their life-patterns even within an environment. Structure and self-organization are the marks of substantiality for Aristotle; that an animal or plant ultimately consists of earth, water, etc. is an inescapably important fact about it, but is not what we should first be attending to when we consider what makes this being real. It should be clear, then, that life is not the exception but the rule in Aristotle's universe. Far from being a sporadic accident, the existence of living beings, immortal celestial ones and mortal terrestrial ones that endlessly recur, is an eternal, irreducible, fundamental fact about the physical world. Moreover, the fact that this world exists and has the nature and kinds of contents it has is not something that Aristotle could have regarded as metaphysically contingent. These considerations help to explain why a model of causation that particularly fits the biological domain becomes, in his hands, a model of causation as such.

I shall now describe some features of causality implied by the purposive-agency model. These have all contributed to historically important arguments.

3.1 'Agency', 'Efficacy', 'Bringing about', 'Production'

These terms are roughly synonymous for the nuclear relation of the Aristotelian efficient

(agent) cause to its effect—what Hume called 'necessary connexion' (A *Treatise of Human Nature* 1. 3. 14; *An Enquiry Concerning Human Understanding* 7).[14] In the history of modern philosophy they denote the object of Hume's empiricist deconstruction, whereby the supposed 'idea' represents only the impression of mental transition that occurs when an experiencing mind flits in expectation from some presented item to the lively idea of its usual concomitant. Since ripples from Hume's critique are still rocking boats to this very day, it is not uninteresting to observe, by way of contrast, the unquestioning serenity with which Aristotle, like common sense, takes it for granted that such terms as 'efficacy' are meaningful, that the corresponding relation (efficacy 'out there') is real and metaphysically unweird, and that propositions predicating it pose no special problem of verification. Aristotle uses, without seeing any need to analyse, the notion of agency or efficacy, in order to spell out one aspect of what he *does* think needs spelling out, namely a scientifically adequate answer to the question 'What is an explanation?' Aristotle's unruffled way with 'efficacy' seems remarkable when we simply compare it with Hume's treatment, but less so when we place it alongside Aristotle's own handling of another of his aspects of scientific explanation, the one labelled 'material cause'. To say that silver is the material cause of the bowl is to say that the information that the bowl is made of silver is indispensable when it comes to explaining the bowl and its formation. Here Aristotle (and all those with whom he would have found himself arguing) simply takes it for granted that we understand and know what it is for something to be made of something. The attention is not on *being made of*—on whether this relation is epistemologically accessible or metaphysically wholesome (Hume, no doubt, should have called these assumptions too into question)—but on whether, or to what extent, *Y*'s being made of *X* ought to enter into the scientific explanation of *Y*

3.2 First Causes

From the modern perspective, one can give a successful causal explanation of *Y* in terms of *X* even if one is unable to show what the cause of *X* is, and so on back. The chain of causes that themselves are effects can be infinite, and we tend to think of it as stretching indefinitely back into the past and forward into the future. It is quite otherwise with the ancient purposive agency model. Here, items in any one causal series are standardly simultaneous or contemporaneous with one another. Moreover, any such series is necessarily finite, and a causal explanation risks being unacceptably incomplete if it fails to connect the initial *explanandum* with the first or ultimate cause.

Before getting to details, we must bear in mind that because this model prescribes a fourfold approach to causal explanation, the question whether a cause has been properly assigned if the cause of the cause (should there be one) has not also been assigned, can theoretically be raised for each of the interlocking aspects, efficient, material, formal, and final.[15] Here I shall focus only on the final cause or end, *and* the efficient cause or *agent.*

It is plausible that an explanation in terms of the specific end *E* for the sake of which some process or formation exists is inadequate if *E* is for the sake of a specific ulterior end, *F,* and *F* is left obscure. For very often the means *M* to a given sort of end *E* can take significantly

different forms, and the fact that E is being implemented for the sake of the ulterior F, rather than for some other possible ulterior, G, is something we need to know if we are to understand properly why M has the specific form that it does. For example, different views can be taken of the ultimate purpose of the circulation of the blood in mammals, and one's understanding of the mechanism of the heart depends on grasping not only that it is for the sake of circulation but also what circulation itself is for. It is perhaps also obvious that an explanation that refers to the action of an agent, B, is unsatisfactory if the action of B is due to the action of an ulterior agent A, and the latter is left obscure. For if B is not itself the ultimate agent, then B is the instrument, or its action the means, by which the ultimate agent brings about some effect. But we do not properly present this effect so as to illuminate it if we relate it only to B, and not also to the agent managing B and this agent's wider purpose. If a hammer were an ultimate agent rather than a tool in the hands of an ulterior being, then the going in of the nails, or the flattening out of the sheet of metal, would be adequately explained by reference just to the action of the hammer; one would be saying in effect: this change is happening because this is what a hammer makes happen, and it does it for no further reason but for its own sake, and because it is the nature of a hammer to behave like that when conditions permit. As things are, we explain the going in of the nails by reference to the hammer-wielding agent: to his or her end in wielding the hammer. Or (as Aristotle often indicates) we explain the going in of the nails by reference to the activity and the end of the *architektōn*, the master-craftsman who guides the manual workers, who stands to them (for purposes of this analysis) in the relation in which they stand to their tools.

The requirement that there be a first, and therefore uncaused, agent of change is part and parcel of the purposive structure of the examples which we have been considering. The key point is that the causal intermediacy of a causally intermediate agent or action (such as the hammer or its action in the last example) consists in its functioning as *instrument* or *means* of a causally prior agent. Although the latter may, of course, itself be functioning as the instrument of an agent yet causally prior to it, we seem intuitively compelled to postulate a stopping point: a causally first agent whose single overall end unifies the entire series. We cannot make sense of a series in which every agent, as we go back causally, turns out to be no more than the means or instrument of some prior agent ad infinitum. The unity of the entire series consists in the fact that the first agent's agency, and the endhood of the first agent's end, are explanatorily relevant to every item in the series. Thus the hammer, or its action, is not to be explained just as the hammering agent's means to a sheet of metal: it is equally to be explained as the non-manual *architektōen*'s means not only to a sheet of metal, but also to the more comprehensive artefact, say a set of armour or a temple with doors of bronze, of which the sheet is a part. It is not the case that the *architektōns* instruction (C_1) gives rise to the muscular activity of the hands-on metalworker, and, as a matter of separate causal fact, it so happens that the metalworker's activity (C_2) gives rise to the movements of the hammer, and, as a matter of yet further causal fact, it so happens that the movements of the hammer (C_3) giverise to the flattening of the metal. On the contrary, C_1 is, as it were, designed precisely to produce the flattening of the metal (and the ultimate artefact) via C_2 and C_3, and this is reflected in the specific form taken by the instruction, C_1.

Since the phrases 'first cause' and 'uncaused cause', with the initial letters tending to be

capitalized, have often been used elliptically for God, or for a unique ultimate cause of the entire universe, it should be pointed out that no such meaning or reference is entailed by the expressions as forged in the original context of Aristotle's purposive-agency model. As our examples indicate, to say that an agent *A* is ultimate or first in a causal series, and is therefore its uncaused cause, is not to imply that *A* is the cause of the entire universe, or that it transcends the order of nature, or operates from outside space and time. Nor is it to imply that *A* operates randomly, indeterministically, or inexplicably. It is simply to say that no ulterior agent is wielding or manipulating *A* for a purpose of its own. This is compatible with holding that *A*'s operating as it does is physically necessary, either absolutely or necessary under the circumstances, given the nature of *A*. It is also compatible with holding that the event is not inexplicable. For in saying that *A* behaves as it does because it is its nature to do so (or to do so under these circumstances), we have not rejected the challenge to explain *A*'s behaviour but have taken the final step in explanation.

There are various motives for refining the notion of cause. Aristotle's was an interest in providing the most informative and illuminating method of explaining the central natural phenomena of his universe. A different sort of motive is created by problems of free will and responsibility,[16] of which readers may have been reminded by the reference to indeterminism in the previous paragraph. The thought that our free and responsible behaviour is caused by factors over which we have no control has often seemed impossible to accept and impossible to reject. The challenge then is to refine the notion of cause either so that the thought becomes more acceptable or so that it becomes more rejectable. Moderns seeking conciliation here have generally taken the first tack, stripping down the notion so that being caused to act is emptied not only of any suggestion of acting under coercion, but also of any suggestion of manipulation by a puppet-master beyond the self, executing purposes of its own. 'Cause' is now understood in terms of the Humeanly inspired regularity theory, or one of its sophisticated (but equally, from the present point of view, unthreatening) descendants. It is now unparadoxical to speak of free and responsible behaviour as a link in a chain of causes stretching back beyond itself. We can see, however, how an account such as Aristotle's can also be turned to use in the dilemma about free will. The move would be to retain his rich conception of the purposive-agent cause while pointing out that free and responsible human agents, just like the natural substances that the scientist studies, function as genuinely uncaused causes in the above sense even though they and their actions are thoroughly embedded in their wider spatio-temporal context, perhaps even operating from it deterministically.[17] We run no risk of denying this when we deny them the status of intermediate links in an Aristotelian causal chain.

3.3 Unmoved Movers

This seemingly mysterious concept, like that of first or uncaused causes, is often assumed to point to something supernatural or theological. But although it, like the other, appears in important theological contexts, its application is not confined to them.[18]

The relevant sense of 'movement' needs explanation. 'Movement' traditionally translates the Greek *kinēsis,* which covers all types of process in which something changes from one

state to another: not merely locomotion but also growth, development, and qualitative alteration. Thus 'moved' and 'mover' here mean the patient and agent of any sort of process. The concept of an *unmoved* mover was developed in conjunction with a sort of contrary: that of a *moved* mover. A moved mover is an agent of change that operates only through being, itself, moved (i.e. through having some process of change occur in it). In our previous example the hammer and, for that matter, the subordinate worker who wields it, are moved movers. Now, Aristotle holds the principles (P_1) that every movement is due to some mover;[19] and (P_2) that no object O can strictly speaking be mover in relation to a movement occurring in O itself.[20] He also assumes that a causal series of movers and objects moved fits the agent-purposive pattern explained above. Consequently, for every such series there is a first mover. And the first mover is unmoved.[21] For being first, it is not moved by anything else; and by P_2 it is not moved by itself. Now, its being unmoved does not mean merely that it is not itself in movement through being moved by some mover. For logically that allows that the first mover might simply be in movement but without being passively mov*ed by* anything: there might be a movement occurring *in* that mover, but not itself due to any mover. However, this possibility is ruled out by Aristotle's P_1. Hence for Aristotle the first and unmoved mover in a series of movers and objects moved is literally without movement: it is unmoved not merely in the sense of not being mov*ed*, i.e. subject to some distinct agent of movement, but in the sense of being without movement at all.

Now this is less radical a notion than it might seem, for if we return for comparison to the concept of a *moved* mover, it is obvious from Aristotle's illustrations that a moved mover for him is something that induces movement in something else through, or on account of, being moved by something itself. (Thus the hammer beating out the metal is a moved mover, and so is the worker who wields the hammer: he is moved and kept moving by the master-craftsman's word.) It follows that an *unmoved* (and, as we have seen, *movementless)* mover is one that induces movement in something else but not through, or on account of, being in a condition of movement itself. Hence what Aristotle means by an unmoved mover is not necessarily something that is absolutely changeless, for an unmoved mover may be in movement as long as its movement, or its being in movement, is causally irrelevant to its functioning as mover of something else. So if the *architektōn* happens to pace round the room while framing his instructions, this does not render him a moved mover in the sense that concerns us if the pacing is causally irrelevant to his craftsmanly efficacy. Causally he is still unmoved and first in the series.

But notwithstanding this clarification, we are probably struggling to comprehend. How can anything induce movement in anything else otherwise than through or by means of some movement that it itself undergoes? The *architektōn* induces movement in his subordinates by issuing instructions, and issuing instructions is causally impossible without moving the body in various ways. Surely a similar tale would have to be told about any natural agent or mover: thus if we are to find a truly unmoved mover anywhere, it seems we must look beyond the natural universe for something divine and mysterious if not also mythological.

Aristotle in his most famous discussions holds that the divine is both a source of physical movement and necessarily absolutely changeless; and without question his concept of 'unmoved mover' is an important building block in this theology. But in itself the concept

also applies within the order of nature. To revert to our example: the agent or mover that we know as the *architektōn* is a complex consisting of human body and soul. Not everything about this individual is essential to its function as mover. But for the purpose of analysis Aristotle tends to identify the mover proper with what is thus essential. So what is the mover proper, the item that is mover and nothing but? It is the 'form in the soul' of the craftsman, by which Aristotle means the craftsman's skill-in-action on a given project.[22] (A skill that is not currently active in connection with a project—its owner may be resting or out of a job—is not the mover of anything, nor does it render the owner a mover in the ordinary, less analytic, sense.[23]) The most convenient way to think of the skill-in-action is as a *definite instruction or set of instructions*. Such an entity does not exist in full definiteness, in the detail needed actually to guide effective behaviour here and now, except when the concrete psychic and physical system in which it becomes, so to speak, 'inscribed', is completely ready to be directed by it; and then the instruction does actually direct, and thereby functions as an effective mover (in the absence of interference). Now this mover that is nothing but a mover is not by itself sufficient for the occurrence of the movement, since that depends on organic and external enabling conditions. Even so, this 'pure' mover is alone responsible for the movement's shape and direction, as distinct from implementation. And arguably the movement's nature and identity (like those of an organic substance) depend more on its shape and direction than on anything else. What, finally, we should notice about the instruction is that precisely by remaining the same and unchanging it performs its function as director of the movement. It stays put until it has been completely carried out. Only by such changelessness-until-completely-implemented can it get completely implemented. This is how this mover cannot function except through being unmoved.

FURTHER READING

Ackrill (1981) and Lear (1988) are useful general introductions to Aristotle. Vlastos (1969; 1971; 1981) is a classic for the history of the development of causal concepts. Sorabji (1980) is invaluable for its penetrating coverage of causality and related topics in Aristotle. Hankinson (1995) and the relevant chapters of Hankinson (1998) give full and clear discussions of Aristotle's approach to science. Cooper (1982; 2004) does much to make sense of Aristotelian teleology. Waterlow (1982) compares Hume and Aristotle on 'efficacy'.

REFERENCES

ACKRILL, J. L. (1981). *Aristotle the Philosopher*. Oxford: Oxford University Press.

COOPER, JOHN M. (1982). 'Aristotle on Natural Teleology' in M. Schofield and M. Nussbaum (eds.), *Language and Logos*. Cambridge: Cambridge University Press, 197-222; repr. in John M. Cooper (2004), *Knowledge, Nature, and the Good*. Princeton: Princeton University Press, 107-29.

HANKINSON, R. J. (1995). 'Philosophy of Science', in J. Barnes (ed.), *The Cambridge Companion to Aristotle*. Cambridge: Cambridge University Press, 109-39.

—— (1998). *Cause and Explanation in Ancient Greek Thought*. Oxford: Oxford

University Press.

LEAR, JONATHAN (1988). *Aristotle, the Desire to Understand*. Cambridge: Cambridge University Press, chs. 1-2.

LONG, A. A., and SEDLEY, D. N. (1987). *The Hellenistic Philosophers,* i. Cambridge: Cambridge University Press.

SORABJI, R. (1980), *Necessity, Cause and Blame: Perspectives on Aristotle's Theory*. London: Duckworth.

VLASTOS, G. (1969). 'Reasons and Causes in the *Phaedo'*, *Philosophical Review* 78: 291-325; repr. in Vlastos (ed.) (1971), *Plato, i. Metaphysics and Epistemology*. New York: Doubleday, 132-66, and in Vlastos (1981), *Platonic Studies*. Princeton: Princeton University Press, 76-110.

WATERLOW [BROADIE], S. (1982). *Nature, Change, and Agency in Aristotle's* Physics, ch. IV ('Agent and Patient'). Oxford: Clarendon, 159-203.

CHAPTER 2
THE MEDIEVALS

JOHN MARENBON

1. INTRODUCTION

Medieval philosophers used the language of cause and effect as frequently as philosophers do now. Viewed very, very broadly, they had in mind a similar notion, at least when they were speaking of efficient causality (and even the three other types of Aristotelian cause—material, formal, and final—can be brought loosely under this concept): causes are in some sense prior to their effects, which they produce and the existence of which they explain. Viewed more closely, medieval notions of causality are sharply different from contemporary ones, and these differences are especially evident in explicit discussions of causation. This short chapter on a vast period—a thousand years of Arabic and Jewish, as well as Latin, culture—can aim to do little more than indicate the variety of these differences, which lie sometimes a little beneath the surface of treatments that seem deceptively approachable. Section 2 will discuss the idea of essential causation—a central theme for Aquinas (1224/5-74) and Duns Scotus (1265/6-1308), which needs to be understood in the light of their main source for it, the Persian philosopher Avicenna (ibn Sīnā, 980-1037). Section 3 looks at the aspect of medieval thought about causation that seems to come closest to the modern debates instigated by Hume, the supposed medieval occasionalists such as the Islamic thinker al-Ghazālī (1058-1111) and the Parisian Arts Master and student of theology, Nicholas of Autrecourt (d. 1369), and the critiques of occasionalism offered by Averroes (c.1126-98), who wrote in Muslim Spain, and Aquinas. Section 4 looks briefly at two discussions of causation from the early medieval Latin tradition, before Aristotle's treatment of the subject was known.

2. ESSENTIAL CAUSATION

In the second of his 'Ways' (arguments for the existence of God: *Summa Theologiae (ST*: Thomas Aquinas 1962) I q. 2 a. 3), Aquinas puts forward the following reasoning. There are efficient causes and effects—so, according to Aquinas, we discover from what we perceive with our senses. Since every efficient cause is prior to its effect, none can be its own cause, because it would then be prior to itself, which is impossible. And, says Aquinas, 'it is not possible to proceed to infinity in efficient causes': without a first cause, there would be no other causes and effects. A first efficient cause must therefore be posited, 'which everyone calls God'.

At first sight, this argument may seem very weak indeed, since there are no obvious grounds for accepting the premiss that there cannot be an infinite chain of efficient causes. Is Aquinas not confusing the genuine requirement that each item in the causal chain has a cause with a spurious one, that there should be a first cause? There is no good reason for him to deny that, for example, the chain of fathers and sons might run on infinitely, since, unlike some of his contemporaries, he did not think that the eternity of the universe could be ruled out on rational grounds. But, as it turns out, he indeed accepts that there might be an infinite chain of fathers and sons. So he explains later on in the *Summa Theologiae* (I q. 46 a. 2 ad 7), answering an objection to his position on the eternity of the world:

It is impossible to proceed to infinity in efficient causes *per se*—that is to say, that the causes required *per se* for some effect are multiplied to infinity; as if a stone were moved by a stick, and stick by a hand, and so on to infinity. But it is not considered impossible to proceed to infinity *per accidens* in agent causes—that is to say if all the causes that are multiplied to infinity are ordered just to one cause and their multiplication is *per accidens;* as a craftsman uses many hammers *per accidens,* because one after another is broken. It is, therefore, an accident of this hammer that it acts after the action of this other hammer. And, similarly, it is an accident of this man, in so far as he generates, that he is generated by another man. For he generates in that he is a man, not in that he is the son of another man. For all men who generate have a position among efficient causes, that is to say, the position of a particular generator. And so it is not impossible that man should be generated by man to infinity. But it would be impossible if the generation of this man depended on that man, and on an elementary body, and on the sun, and so on to infinity.

The distinction made here between essential (*per se*) and accidental (*per accidens)* causation seems to give Aquinas what he needs to help this causal argument for God's existence, but in what exactly does the distinction consist? The passage suggests that, in an accidental chain, each cause is itself regarded as an independent causer of its effect: although there could be no grandson without the grandfather's engendering of the father, the father exercised this function independently of his own father. In an essential chain, the subsequent causes are dependent—the hand, stick, and stone example suggests entirely dependent—on the first cause for their own causal efficacy. Beyond this, it is difficult to gloss the passage further without looking more widely. One obvious text to turn to for illumination is the *Metaphysics* from Avicenna's *Shifā* (Cure)(= *M)*, which Aquinas knew in Latin translation, since it is from there that the distinction between these types of causes is taken.

Avicenna was a follower of Aristotle, but he read him in the light of a long tradition of Peripatetic and Neoplatonic commentary, and he aimed to combine and adapt his materials into what he believed was true doctrine, rather than to interpret ancient texts literally. One of his innovations was to distinguish between a physical and metaphysical sense of efficient causation (*M* VI. 1. 2-3; Avicenna 2005: 194-5). Physics, he goes on to explain, examines causes of motion, and such causes are temporally prior to their effects. But metaphysics examines the causes of existence, which are simultaneous with their effects. Avicenna understood causes of existence in the context of the theory of cosmological emanation he took over from the Neoplatonists, as adapted by his Islamic predecessor, al-Fārābī. From God there

emanates a First Intellect, which in its turn produces the celestial sphere that it moves and a Second Intellect. This process continues until it reaches the tenth and lowest Intellect, the Active Intellect, which does not have its own celestial sphere but, rather, produces the forms and matter of the sub-lunar world. Although not just God, but all the Intellects and their spheres are, he believes, eternal, God is necessary in himself. Avicenna explains this idea of necessity in terms of causality. That which is necessary in itself has no cause; everything else, which is not necessary in itself but merely possible, owes its existence to a cause; given the cause, it is necessary too, but through another, not in itself. Avicenna gives demonstrations to show that there can be only one being that is necessary in itself, through which all other beings —even those that are eternal—exist necessarily (*M* I. 6–7; Avicenna 2005: 29–38; cf. Marmura 1984).

Aquinas's essentially ordered causes seem to be very like the causes Avicenna's metaphysician investigates. The closeness is especially evident in the Third Way, the argument for the existence of God that immediately follows the one discussed above, where Aquinas adds to a line of reasoning close to the Second Way a distinction on openly Avicennian lines between things that have the cause of their necessity from elsewhere and that which has the cause of its necessity in itself. None the less, Aquinas is using Avicennian ideas to give a very different view of how divine causality operates and of causation itself. Aquinas entirely rejects Avicenna's idea of a chain of causation in which the first cause is directly responsible only for bringing about the First Intellect, and each subsequent level of being is directly caused by the one above it. Unlike Avicenna, he makes a very sharp distinction between ordinary causation and the special type of causation he calls 'creation', and he maintains that it is both heretical and contrary to reason to attribute the power to create to anything other than God—which he takes Avicenna's theory to do (*De potentia* q. 3 a. 4 (Thomas Aquinas 1965, 1932-4); *ST* I q. 5 a. 5). But what does he mean by 'creation'?

He defines it (*ST* I q. 45 a. 1) as the 'emanation of the whole of what exists *(totius entis)* from the universal cause'. But he does not mean to present it just as the first cause that initiates the rest of an essentially ordered causal chain. Rather, Aquinas thinks (Commentary on *Metaphysics* VI lec. 3 (Thomas Aquinas 1971); cf. *ST* I q. 44 a. 2) of causation as taking place on three levels. The first level is particular causation of a member of a species by a member of the same species (for instance, a son by a father): it is such causes which, if they themselves are put into a chronological chain, could in principle stretch on infinitely. A second level is the causality exercised by the heavenly bodies. Aquinas considers that the way in which things on earth (but not human minds) are disposed is largely, though not wholly, determined by the movements of the sun and the stars, and that different chains of particular causes are caused by a single such more universal celestial cause. The second level of causality thus explains the existence and way of existence of the same things as particular causes explain, but at a higher level of generality. But there is an even higher level of generality. All things have it in common that they are existing things, *entia*, and this is brought about by the highest cause of all, God.

Aquinas's conception of God as the cause of existence (*esse*) rests on another idea taken from Avicenna, but transformed in its use. In everything other than God, Aquinas considers (*ST* Iq. 3 a. 4), a thing's essence—its being a thing of this sort or that, an angel, a human, or a stone, for instance—can be distinguished from its existence. From the fact that *A* is, say, a

human, it does not follow that *A* in fact exists. But in the case of God, the First Cause, his essence is *esse*—to exist: were it not, God's existence could be explained only by its being caused by something else, and so he would not be the First Cause. Since, then, God alone is essentially existence, the existence of all other things is by participation in his existence (cf. e.g. *Summa contra Gentiles* lib. 3 d. 66 n. 7; Thomas Aquinas 1961, 1975). Aquinas quotes the Neoplatonic dictum that the higher the cause, the more widely its effects extend (*ST* Iq. 65 a. 3). It is true of every thing, and so of every effect, that it exists, and so it is right that the First Cause should be the cause of existence.

For Avicenna, then, metaphysical causation—which is sharply distinguished from the physical causality of motion—concerns the giving of existence, which is passed down through a causal chain from God, the one existent necessary of itself. For Aquinas, the causality of all agents other than God is exercised to bring about *how* things are, but *that* this or that thing exists is the direct causal effect of God.

Aquinas's conception of God as the necessary First Cause in an essentially ordered causal chain immediately raises the problem of how there is room for contingency in the world. Aquinas believes that secondary causes do not take on the necessity of their first cause. He gives the example of the flowering of a plant. Its first cause is the sun, and the secondary cause is its own capacity for growth. Although the sun's motion is unvarying, a defect in the plant's capacity for growth can prevent the flowering from taking place (*ST* I q.14 a.13 ad 1).

This answer was rejected at the turn of the fourteenth century by Duns Scotus *(Lectura* I. 39; John Duns Scotus 1994: 90-4). He has two main arguments. First, in an essentially ordered causal chain, a secondary cause moves only in so far as it is moved by the first cause. If it is moved necessarily, then it will bring about its own effect necessarily too. Second, take the causal chain $A \rightarrow B \rightarrow C$, and suppose that—as Aquinas holds—*A* causes necessarily and *B* causes contingently. If *C* received its being first (in some perhaps logical, rather than temporal, sense) from *A* necessarily, and then from *B* contingently, it would be both necessary and contingent; since that is impossible, *C* cannot receive contingent being from *B*. Suppose, though, that *C* receives its being simultaneously from *A* and *B*, would *B* then make it contingent? No, answers Scotus, because if *C* has a necessary relation to *A*, then even if *B* did not exist, *C* would exist, and so it does not gain its being from the contingent cause *B*. For these reasons, Scotus holds that, in order to explain the contingency that exists in the world, we must say that the First Cause itself causes contingently. Scotus's own argument for the existence of God (proposed both in his *De primo principio* (John Duns Scotus 1966) and in his *Sentences* commentary *(Ordinatio* I d. 2 q. 1; John Duns Scotus 1987: 35-81) is based on a systematic presentation of the nature of essentially ordered causal chains. In this respect, it resembles Aquinas's First, Second, and Third Ways. But Scotus's theory of God as a *contingent* cause (cf. Sylwanowicz 1996) colours his thinking in a way that brings it even further away from Avicenna's conception of causes ultimately dependent on a necessary cause.

3. Causation and Occasionalism

The passages just examined make many, varied uses of ideas about causality, but it may be questioned whether they offer an *analysis* of causation. And, as signalled by their authors'

strong tendency to regard causes and effects as things (see below, sect. 4), their central concern is, in any case, creation, a very special type of causality that plays little part in more recent discussions. Medieval occasionalism offers the prospect of a discussion that links a little better with modern questions.

The most influential medieval occasionalism was linked to *kalām,* a type of philosophizing in Islam that, unlike Avicenna's, took place within the context of Muslim theology. The most argumentative and intellectually adventurous school of early Muslim theologians, the Muʿtazilites, borrowed (probably indirectly) various motifs from Aristotelian and Neoplatonic thought and shaped them into a variety of non-Aristotelian physical and metaphysical theories. Many of the Muʿtazilites were atomists, some of whom took humans and other objects in the world to be mere conglomerations of accidents, which persist as apparently stable entitites because God wills their perdurance from moment to moment. Al-Ashʿarī (d. 935), trained as a Muʿtazilite but founder of his own, more traditionalist school of theology, brought together occasionalist elements in earlier *kalām* into a comprehensive occasionalist position. Accentuating the Muʿtazilite view that things do not have natures that guarantee some constancy in their behaviour, but are made up of accidents, Ashʿarī claimed that, when accident b precedes, accompanies, or follows accident a, it could always be the case that God might have created an accident b^*, opposite to b, to precede, accompany, or follow a. Moreover, he denied that anything, animate or inanimate, acts: all apparent action in the world is merely borrowed from the one true agent, God (Perler and Rudolph 2000: 51–6).

Ashʿarī merely presents a position—one that fits well into his general theological outlook. But, in the next century, al-Ghazālī (1058–1111), probably the most influential of all Muslim theologians, would give a remarkable analysis of the arguments for and against occasionalism. His most famous treatment of the issue is in his *Tahāfut al-falāsifa,* 'The Incoherence of the Philosophers', Discussion 17. The title of this treatise promises an attack on the 'philosophers'—that is to say, on the Aristotelian-Neoplatonic tradition in Islam, which was represented by, above all, Avicenna. But little is straightforward about Ghazālī, and, while rejecting some parts of Avicenna's thought, he ended by introducing much of it into the tradition of *kalām.*

The two-sidedness of Ghazālī's approach is nowhere more evident than in his discussion of causation. His reason for raising the issue is that the philosophers' approach to natural science excludes the possibility of miracles. Ghazālī wishes to accept as literally true the traditional miracles of Muslim belief, and the ultimate aim of his analysis is to provide a way, or ways, to do so that are as well supported rationally as the philosophers' contrary position. He begins with the statement of a view (call it '*V*') that sounds like a manifesto of occasionalism:

The connection between what is habitually believed to be a cause and what is habitually believed to be an effect is not necessary, according to us. But with any two things, where this is not that and that is not this, and where neither the affirmation of the one entails the affirmation of the other nor the negation of the one entails the negation of the other, it is not a necessity of the existence of the one that the other should exist, and it is not a necessity of the non-existence of the one that the other should not exist: for example, the quenching of thirst and drinking, satiety and eating, burning and contact with fire, light and the appearance of the

sun, death and decapitation, healing and the drinking of medicine ... Their connection is due to the prior decree of God, who creates them side by side, not to its being necessary of itself, incapable of separation. (al-Ghazālī 2000: 166: Marmura's translation, punctuation amended slightly)

As the Discussion has often been read, *V* is juxtaposed with a more Aristotelian view (§18; al-Ghazālī 2000: 171–3): the interpretative question is then thought to be whether Ghazālī is giving up the occasionalist theory with which he supposedly began (Courtenay 1973; cf. Frank 1994), or whether his underlying position remains that of the occasionalists (Marmura 2005: 145–53). A more sophisticated examination of the literary structure of the Discussion (Perler and Rudolph 2000: 63–105) leads to the conclusion that Ghazālī is trying to devise a theory that would satisfy both the philosophers and the Ash'arite theologians. But the truth is rather that Ghazālī does not produce a compromise, but rather reframes the terms of the dispute in a way that accommodates and goes beyond both earlier positions.

Ghazālī's procedure is to examine, and criticize or defend, a variety of theories about causation. Theory *A* is straightforward Ash'arite occasionalism (§§5–6; al-Ghazālī 2000: 167–8). Suppose a piece of cotton is brought into contact with a fire: it is wrong, on this theory, to say that the fire causes the cotton to burn. The causal agent is God, directly or through his angels, who turns the cotton into black cinders when the fire, which itself has no action, touches it. The argument given for *A* is that the opponent has no proof that fire is the agent: all he can observe is that, at the time it comes into contact with the fire, the cotton burns, not that the fire burns it. This line of thought is bolstered by a thought experiment. Imagine someone blind from birth suddenly cured. He would immediately jump to the conclusion that it is his restored power of sight that enables him to see the world and its colours; but, come the evening, he would realize that it is in fact the sunlight that is the cause of his seeing the colours. Ghazālī accepts, however, that *A* leads to some alarming consequences (§13; ibid. 169–71). If every event is caused by God, who can act as he wills, then none of the regularities in nature which we take for granted can be relied on: for all I know, there is a lion about to pounce on me as I type these words, but God has not created in me the sight of it. Moreover, anything can change into anything:

I do not know what is at the house at present. All I know is that I left a book in the house, which perhaps now is a horse that has defiled the library with its urine and its dung, and that I have left in the house a jar of water, which may well have turned into an apple tree. (Marmura's translation)

The relish with which Ghazālī's develops these consequences suggests that he does not think that, in its straightforward form, *A* is salvageable. But there is more sophisticated form of the theory (*AA*) that he does defend. *AA* (§§15–17; al-Ghazālī 2000: 170–1) adds to *A* the idea that God gives us—simply, so it is suggested, through habituation—the knowledge that the world will follow the regular order it has always done, even though it remains a possibility that God *could* bring about any of the irregularities the critics of *A* mention. The only times

when God disrupts the regular order of things is when he produces a miracle, and God does not let the knowledge of these miracles disrupt people's expectations that things will continue their regular course.

AA is left by Ghaz**ā**l**ī** as a tenable theory, but he spends more time discussing the ways in which the philosophers' position can be refined. The simple philosophical theory, *P*, is that things in the world, such as fire, have a nature that makes them necessarily act in a certain given way under the appropriate conditions: fire burns whenever it is in contact with something burnable (§4; ibid. 167). Ghaz**ā**l**ī** does not spend time on *P* because, in his view, it is not what is held by the most sophisticated philosophers. According to their (broadly Avicennian) theory, *PP* (§§8–9; ibid. 168–9), the cause for events on earth is the Agent Intellect, which emanates forms; but there are observed causes here on earth that prepare for the reception of these forms. *PP* allows no more place for miraculous exceptions to natural regularities than *P*. Since Ghaz**ā**l**ī** is starting from the position that the philosophers are wrong in so far as they go against accepted Islamic beliefs, *PP* is unacceptable to him. But he develops, and raises no objections to, a modified, elaborated version of the theory, *PPP* (§§18–25; ibid. 171–3). *PPP* accepts the structure of *PP*—a heavenly first cause (Ghaz**ā**l**ī** envisages God and his angels, rather than the Agent Intellect)—with an input into the causal process from the natures of earthly things, which determine whether or not they are disposed to receive the forms that emanate from above. Unlike the first cause of *PP*, the first cause of *PPP* does not operate according to necessity but according to will. There is no need, however, Ghaz**ā**l**ī** suggests, that in acting as cause God need fundamentally go against the natures of things; rather, he makes small, temporary changes (a prophet is thrown into the fire and not burnt—how? The fire retains its power to burn, but it is temporarily confined so that it does not burn anything else) or God speeds up natural processes.

Towards the end of the Discussion (§§27–39; ibid. 174–7), the argument takes another turn. Ghaz**ā**l**ī** introduces a philosopher into the discussion, who begins, in conciliatory tones, by saying that thinkers like him 'help' the theologians, by holding that whatever is possible is within God's power, whilst the theologians correspondingly help the philosophers by accepting that God cannot do what is impossible. The question is, therefore, what is impossible? The philosopher refers back to the position announced by Ghaz**ā**l**ī** at the beginning of the passage, where only what is logically contradictory is held to be impossible. Such a narrow definition of impossibility, he contends, leads to unacceptable consequences: God could, for instance, give knowledge to something inanimate, or he could change an accident into a substance. He could make a dead man perform all the actions of a living one, thereby confounding the distinction between voluntary actions and involuntary movements, since the purposiveness of this dead man's behaviour would indicate, wrongly, that it was voluntary, the result of the agent's will and power. The unstated point behind the philosopher's comments is that the theologians need to admit that it is also impossible for there to be a cause without its having its effect.

Ghaz**ā**l**ī** does not accept this conclusion, because he considers that the philosopher's objections can be answered without broadening the definition of impossibility, but by being more attentive to what, when analysed, entails a contradiction. It is, he contends, genuinely impossible for God to give knowledge to something inanimate, because to be inanimate means

to lack apprehension, and therefore knowledge. If God gave something knowledge, it would no longer be inanimate. Similarly, when a thing changes from one genus to another, what happens is that the same matter takes on one form and then another. Since substances and accidents have no common matter, it is impossible for one to be changed into the other. Giving a dead person the actions of a living one is not, however, impossible, although we are habituated to taking such actions to indicate a live agent. But we would still be able to reason from the purposiveness of actions to a knowing, and so voluntary, agent—not, in this case, the human, but God.

What does Ghazālī hope to achieve by raising these questions here and answering them in this way? He is both rejecting a central tenet of the philosophers' position—that it is impossible for causes not to bring about their effects (when the matter is properly disposed to receive them)—and yet saying nothing to endorse the *kalām* occasionalist view that God alone acts. But he is not proposing some compromise position. He does not go back on what he has established already: that there is both an adapted occasionalist theory (*AA*) and a modified philosophical one (*PPP*) which are each acceptable in their main lines, since neither gives rise to absurd consequences nor removes the possibility of miracles. Rather, in this final section he returns to *V*, the point from which he began and which he regards as central. God can do only what is possible, but causal connections, however understood, do not determine what is possible or not. Either *AA* or *PPP* can be held, so long as it is clear that they do not violate this principle (*AA* can be strengthened, since Ghazālī has provided a principled way of avoiding some of its more extreme consequences with regard to the changing of one thing into another; in *PPP* the power of the first cause to act in ways out of the ordinary will be underlined). Ghazālī gives no judgement as to which of the two theories is in fact correct—and there is no reason why he should, given that the aim of his book is merely to refute philosophical teachings in so far as they contradict Islam.

Ghazālī's *Incoherence* is the subject of a section-by-section critique in the *Tahāfut al-Tahāfut* (*The Incoherence of the Incoherence*) by Averroes. Averroes held a near fanatical admiration for Aristotle, upon whose works he spent much of his career commenting, whilst in his *Tahāfut* his aim is specifically to defend philosophers against the criticism that they contradict Islamic orthodoxy. These two factors help to explain why his discussion of causality does not properly engage with the his opponent's arguments. On the one hand, Averroes insists that the occasionalists are obviously wrong because it is 'self-evident that things have essences and attributes which determine the special functions of each thing' and that all events have the four (Aristotelian) causes. That there are causes and effects is, he says, a matter of logic (Averroes 1930: 520–2; trans. Van Den Bergh in Averroes 1954: 318–19). On the other hand, Averroes reads Ghazālī as proposing *PPP* as his own theory and abandoning any version of occasionalism, and he does not take the final part of the Discussion (§§ 27–39) as marking a new approach. By this strategy he is able, in line with his overall aims, to minimize the difference between his own position and that of the theologians, as represented by Ghazālī (rather than as represented by the Ash'arite occasionalists). He does not think that *this* theological theory need be rejected, but he also proposes a philosophical one (Averroes 1930: 538–42; 1954: 331–3), in which the causal powers of different things are much more tightly determined by their natures than according to the the theological view (on the whole topic, see Kogan 1985).

Averroes's *Tahāfut* was not translated into Latin until 1328 and was influential more on Renaissance than medieval Latin philosophy (Averroes 1961: 24–50). Thirteenth-century and early fourteenth-century Latin thinkers were acquainted with Ash'arite occasionalism through the account of it given by the great Jewish philosopher Moses Maimonides (1138–1204) in his *Guide of the Perplexed,* which had been translated into Latin in the 1220s. There Maimonides summarizes (I. 73; Maimonides 1963: 194–14, esp. 201–3) the theories of the *kalām* theology known to him—predominantly Ash'arite. Aquinas includes his criticisms of occasionalism in a chapter of his *Summa contra Gentiles* (III. 69) labelled 'Against those who deprive natural things of the actions that belong to them'. Among his varied arguments (cf. Perler and Rudolph 2000: 131–45) is the Aristotelian view—close to Averroes's—that different sorts of things have their own characteristics and capabilities that determine how they act, and that, without being able to infer causes from effects causes, we would be deprived of all knowledge of nature. Aquinas also believes that, if he had made his creatures unable to act, the creator would have deprived them of what is in fact best in them, in a way incompatible with God's bounty and omnipotence. Yet Aquinas has to balance his defence of the causality of animate and inanimate creatures with his view (see above, sect. 2) that, in a certain way, God causes all things. He ends his chapter with the comment: 'We do not therefore deprive created things of the actions that belong to them, although we attribute all the effects of created things to God, as it were working in all things.'

Medieval Latin thinkers were usually, like Aquinas, critical of occasionalism. William of Ockham (*c*.1288–1347) was once wrongly presented as a precursor of Hume, but was in fact, like Averroes, an exponent of an idea of causal powers that Hume would have ridiculed (Adams 1987: 741–98; Robert 2002). But one Parisian Master who has often been seen as an advocate of occasionalism is Nicholas of Autrecourt, whose career was ended when his doctrines were condemned in 1346. In his exchange of letters with a Franciscan theologian, Bernard of Arezzo, Nicholas holds that all evident (evidently—that is, without any doubt—certain) propositions are founded on the principle of non-contradiction: 'Contradictories cannot be simultaneously true' (Letter to Bernard II. 2; Nicholas of Autrecourt 1994: 58–9). If I hold that *p,* but it is possible without any contradiction for it to be the case that not-*p,* then it is not evident that *p.* Applying this principle to 'if ... then ... ' propositions, Nicholas claims that they are evidently true if and only if it is impossible that the antecedent is true and the consequent false, *and* the consequent is the same in reality as the antecedent or the same as part of what the antecedent signifies (Letter to Bernard II. 5 and 10; Nicholas of Autrecourt 1994: 60–1, 64–5). Nicholas does therefore accept, not merely obvious tautologies, but propositions such as 'If there is a house, there is a wall' as evident, because a wall is part of what is meant by 'a house' (Letter to Bernard II. 21; Nicholas of Autrecourt 1994: 70–3). But 'if ... then ... ' propositions in which the antecedent and consequent are related as cause and effect will not meet the criteria for evidentness.

In another work (*Exigit ordo* Tr. 1; Nicholas of Autrecourt 1939: 237, lines 39–47), Nicholas explains, without calling on this logical criterion, why causal rules (such as 'the magnet attracts iron') are merely the results of an inbuilt habit to consider as a fixed rule what has happened frequently. Consider the argument

(1) There is *A.*
(2) *A* is the natural cause of *B.*

So:

(3) There is *B*.

Nicholas seems willing, in this work, to accept such an argument as valid. But what is a natural cause? Is it not 'what has produced something in many cases in the past, and will produce it in the future if it lasts and is applied'? If so, then we cannot know that (2) is true, even though *B* has in the past followed *A* in many cases, since it is 'not certain that it should be so in the future'.

Does any of this make Nicholas into an occasionalist? It has been pointed out that, although he shares with the occasionalists the negative thesis that, from the observation of nature, we cannot conclude that there exists a causal connection, he does not share their positive view that God is the sole cause of everything (Perler and Rudolph 2000: 178–83). Moreover, Nicholas is *not* in fact trying to deny that there are causal connections—he even works out an elaborate atomistic theory of causality (Grellard 2002: 277–89). He merely denies that we know *evidently* the relation between cause and effect. Still, there is an uncanny similarity—almost certainly not the result of any direct influence—between Nicholas's position in his letters to Bernard and Ghazālī's final and underlying view, which is compatible with occasionalism or adapted Aristotelianism. As in Ghazālī, causation is shown not to be a matter of logical necessity, and nothing is accepted as truly necessary or impossible unless it is logically so.

The critique of occasionalism provides the setting for the closest medieval analyses of causation. The intellectual agility of Ghazālī or Nicholas of Autrecourt in this area is impressive, but they, like the other contributors to the debate, are working within very particular contexts: so, for Ghazālī, the need to make room for miracles; for Averroes, the defence of Aristotle; for Nicholas, the wish to distinguish evident from non-evident knowledge, as part of a critique of Aristotelianism. As a result, it is hard, except in a very vague and general way, to make connections between these discussions and modern or contemporary treatments of causation.

4. THE VARIETY OF MEDIEVAL DISCUSSIONS OF CAUSATION: THE EARLY MIDDLE AGES

From the discussion so far, sophisticated medieval discussions of causation seem to have taken their starting point from Aristotle, often read through a Neoplatonic lens; even Ghazālī focuses mainly on the Neoplatonic Aristotelianism of Avicenna. Before the early thirteenth century, in the Latin West none of Aristotle was known except for his logic, and yet thinkers still discussed causation, often in a sophisticated way. One example is the ninth-century thinker, John Scottus Eriugena, who, drawing on Neoplatonic thinking through his reading of the Greek Christian writers, distinguishes three types of cause: God, who is the first and the final cause of all things; the 'primordial causes'—a sort of cross between Platonic Ideas and Stoic seminal reasons, which are created by God and are then responsible for the production of the rest of created nature; and what he calls *ousia* (his transliteration of the Greek word for 'being'), which he considers to stand in a quasi-causal relation to the substances that populate

the world (cf. Erismann 2002).

One of the most remarkable pre-thirteenth-century treatments of causation, however, is far more independent of ancient traditions. In his famous discussion of universals in the *Logica Ingredientibus* (*c*.1118–19), Abelard has to explain the semantics of universal words, given that, according to him, there are no universal things. Part of his explanation consists in saying that a universal word is imposed on the objects it signifies 'by a common cause'. Singular humans, for example, although they are each distinct from one another, come together 'in that they are humans'. By this Abelard does not mean, he stresses, that there is any sort of universal human in which they share. They come together in *being* humans, and *to be a human* is not any sort of thing (and, indeed, he observes, singular things can equally well come together in what they are not—for example, in not being white, or not being a human). But then, asks Abelard, how can it be that things come together in what is not a thing at all? It is here that he turns his attention from semantics to the nature of causation: 'Often we give the name of causes even to what are not any sort of things, as when we say: "He was whipped because he did not wish to go to the forum". He did not wish to go to the forum, which is given as the cause, is not a thing (*nulla est essentia*)' (Abelard 1919–33: 20–1).

Most medieval discussion about causation in the medieval Aristotelian and Neoplatonic traditions conceives causes as things—in most cases, corporeal (a father) or incorporeal (the Agent Intellect) substances. The effects of causes are often things too (for instance, the father is the cause of his child), although causes of motion are also discussed. Presentations and critiques of occasionalism may seem to be talking in terms of states of affairs (bringing the fire into contact with the fabric, the fabric burning), but the effects were probably thought of by the *kalām* theologians as accidents, whilst the cause was, of course, on their view, God. All this area, however, is given little attention—the modern reader is left to surmise the ontological status of causes and effects, and to wonder at why the medieval authors left it so vague. By contrast, Abelard addresses the problem explicitly and is happy to allow a *status* (being-a-human) or, in his example about not going to the forum, what he would call a *dictum* (something near to a proposition in the modern sense (Marenbon 2006: 343–4 for an overview)) to be a cause. And, according to Abelard, neither a *status* nor a *dictum* is any sort of thing at all—a point he can underline by using a negative proposition in his example, although the *dictum* of the sentence 'He wished to go to the forum' would equally for Abelard not be a thing at all. Whether, in his thought as a whole, Abelard fully justifies his reductionist position (about, for example, universals and dicta), is still debated; none the less, his application of it to the problem of causation is a striking example of his originality, as well as the richness of the medieval discussion.

FURTHER READING

Esposito and Porro (2002) collect a number of valuable new studies on causation in medieval philosophy. The best general discussion, but limited to occasionalism, is Perler and Rudolph (2000). Important contributions to the debate on Avicenna's attitude to causality are Marmura (1984) and Frank (1994), and on Averroes and causality there is Kogan's very detailed study (1985). There is no good, general study of causation in medieval philosophy.

References

ADAMS, M.McCORD (1987). *William Ockham*. Notre Dame, Ind.: University of Notre Dame Press.

AL-GHAZĀLĪ (2000). *The Incoherence of the Philosophers*, trans. (with parallel Arabic text) M. E. Marmura. Provo, Ut.: Brigham Young University Press.

AVERROES (1930). *Tahafot at-Tahafot*, ed. M. Bouyges. Bibliotheca Arabica Scholasticorum, Arabic series 3. Beirut: Imprimerie catholique.

—— (1954). *Tahafut al-Tahafut*, trans. S. Van Den Bergh. Cambridge: Gibb Memorial Trust.

—— (1961). *Averroes' 'Destructio Destructionum Philosophiae Algazelis' in the Latin Version of Calo Calonymos*, ed. B. H. Zedler. Milwaukee: Marquette University Press.

AVICENNA (2005). *The Metaphysics of 'The Healing'*, trans. (with parallel Arabic text) M. E. Marmura. Provo, Ut.: Brigham Young University Press.

COURTENAY, W. J. (1973). 'The Critique on Natural Causality in the Mutakallimun and Nominalism'. *Harvard Theological Review* 66: 77–94

ERISMANN, C. (2002). '"Causa essentialis". De la cause comme principe dans la métaphysique de Jean Scot Erigène' in Esposito and Porro 2002: 187–215.

ESPOSITO, C. and PORRO, P. (eds.) (2002). *La causalità = Quaestio (Annuario di storia della metafisica)*, 2.

FRANK, R.M. (1994). *Al-Ghazālī and the Ash\carite School*. Durham, NC: Duke University Press.

GRELLARD, C. (2002). 'Le statut de la causalité chez Nicholas d'Autrécourt' in Esposito and Porro 2002: 267–89.

JOHN DUNS SCOTUS (1966). *A Treatise on God as First Principle*, trans. (with parallel Latin text) A. B. Wolter. Chicago: Franciscan Herald.

—— (1987). *Philosophical Writings*, trans. A. B. Wolter. Indianapolis: Hackett.

—— (1994). *Contingency and Freedom. Lectura I* 39, trans. with commentary A. Vos Jaczn, A. Veldhuis, A. H. Looman-Graaskamp, E. Dekker, and N. W. den Bok. New Synthese Historical Library 42. Dordrecht: Kluwer.

KOGAN, B. S. (1985). *Averroes and the Metaphysics of Causation*. Albany, NY: State University of New York Press.

MAIMONIDES, MOSES (1963). *The Guide of the Perplexed*, trans. S. Pines. Chicago: University of Chicago Press.

MARENBON, J. (2006). 'The Rediscovery of Peter Abelard's Philosophy'. *Journal of the History of Philosophy* 44: 331–51.

MARMURA, M. E. (1984). 'The Metaphysics of Efficient Causality in Avicenna (Ibn Sina)' in M. E. Marmura (ed.), *Islamic Theology and Philosophy: Studies in Honor of George F. Hourani*. Albany: State University of New York Press, 172–87.

—— (2005). 'Al-Ghazālī', in P. Adamson and R. C. Taylor (eds.), *The Cambridge Companion to Arabic Philosophy*. Cambridge: Cambridge University Press, 137–54.

NICHOLAS OF AUTRECOURT (1939). *Tractatus Universalis* ('Exigit Ordo'), ed. J. R.

O'Donnell. *Mediaeval Studies* 1:179–280.

—— (1994). *His Correspondence with Master Giles and Bernard of Arezzo*, ed. L. M. De Rijk (with parallel English trans.), Studien und Texte zur Geistesgeschichte des Mittelalters 42. Leiden: Brill.

PERLER, D., and RUDOLPH, U. (2000). *Occasionalismus: Theorien der Kausalität im arabisch-islamischen und im europäischen Denken*. Abhandlungen der Akademie der Wissenschaften in Göttingen, phil.-hist. Kl. 235. Series 3. Göttingen: Vandenhoeck & Ruprecht.

PETER ABELARD (1919–33). *Peter Abaelards philosophische Schriften*, ed. B. Geyer. Beiträge zur Geschichte der Philosophie und Theologie des Mittelalters 21. Münster: Aschendorff.

ROBERT, A. (2002). 'L'explication causale selon Guillaume d'Ockham', in Esposito and Porro 2002: 239–65.

SYLWANOWICZ, M. (1996). *Contingent Causality and the Foundations of Duns Scotus' Metaphysics*. Studien und Texte zur Geistesgechichte des Mittelalters 51. Leiden: Brill.

THOMAS AQUINAS (1932–4). *On the Power of God*, trans. Dominicans of the English Province. London: Burns, Oates & Washbourne.

—— (1961). *Liber de veritate catholicae Fidei contra errores infidelium seu Summa contra Gentiles*, ed. P, Marc, C. Pera, and P. Caramello. Turin: Marietti.

—— (1962). *Summa theologiae*: Latin text and English translation, introductions, notes, appendices, and glossaries, general ed. T. Gilby. London: Blackfriars.

—— (1965). *Quaestiones disputatae* II, ed. P. M. Pession. Turin: Marietti.

—— (1971). *In duodecim libros Metaphysicorum Aristotelis* expositio, ed. M. R. Cathala and R. M. Spiazzi. 2nd edn. Turin: Marietti.

—— (1975). *Summa contra Gentiles*, trans. A. C. Pegis et al. Notre Dame, Ind.: University of Notre Dame Press.

CHAPTER 3
THE EARLY MODERNS

KENNETH CLATTERBAUGH

1. THE CHALLENGE: TOO MANY CAUSES

The early moderns confronted an abundance of causes and types of causal explanations. From Aristotle through the Scholastics they had inherited the doctrine of the four causes, that is, depending upon context the material, formal, final, or efficient cause provides the proper explanation or answer to a 'Why?' question. From the Christianization of Aristotle's Unmoved Mover they had inherited the idea of God as creator of the universe. And to the creator role had been added the idea of providence whereby God in some sense 'manages' the world of mundane events (Collins 1960). Aquinas had actually identified God as the efficient cause of all things. Both Aristotle and the Scholastics had sometimes identified the middle term of a scientific syllogism as a cause thereby both blurring the distinction between concepts on the one hand and the entities that they stand for on the other and adding yet another candidate to the list of causally explanatory entities (Clatterbaugh 1999: 11–14).

The pre-moderns had identified both material and immaterial entities as causes to be used in scientific explanation. Neither Aristotle who explained much through the activity of the soul nor the Scholastics who invoked such immaterial entities as substantial forms, the soul, or explanatory qualities such as *moist* and *dry* were committed to any form of naturalistic materialism. Yet, the new materialism of late sixteenth- and early seventeenth-century science challenged the usefulness of immaterial causes. Instead, the new science appealed to successful experimentation and manipulation of the material world as a way to identify causes. Materialism and its frequent companion, mechanism, only added to the complexity of the moderns' philosophical/scientific picture of causation.

A conscientious, philosophically astute, and religious scientist of the early seventeenth century whose intuition was that the best explanation is a *complete* account of the *real* causes of an artefact such as a building would be forced to offer up a complex compendium of causes that would not easily fit together. And even having listed the causes of this artefact, this scientist/philosopher would face the knotty problem of the relative 'strengths' of these causes as well as the 'ranges' of their effects. What, for example, was the role of God relative to the builder relative to the planner? Aquinas (1945: 30), for example, tried to reconcile such multiplicities of causes by identifying both God and the builder as efficient causes that 'concur' in the production of the building. Philosopher/scientists were thus led to celebrate some causes as *primary* while others were regarded as *secondary*; some causes were *general* and others *particular*. But such strategies only moved causes around and did little to simplify

causal explanation.

What was needed, although it was not immediately apparent to the moderns, was a good house cleaning. The number and kinds of causes that were to be taken seriously in proper explanations needed to be greatly reduced and simplified. In sum, that simplification is the story of the early moderns before Hume. The details of that story and the arguments by which the many candidates are either discarded or marginalized is the story to be told in this chapter.

2. DESCARTES

Descartes is hardly the first modern to engage in the necessary house cleaning, that process was well underway under the direction of Francis Bacon in the *New Organon*. Bacon comes close to embracing a manipulative, hence materialistic, theory of causation when he suggests that 'where the cause is not known the effect cannot be reproduced' (1889: 3). In the same work Bacon attacks the very vocabulary by which the Scholastics articulate the concept of causation. He finds among the worthless 'dogmas of philosophies' the syllogism, the concepts of substance, matter, and form as well as the various explanatory qualities such as *moist*, *dry*, *hot* that were used by scholastic science.

But Bacon is more the scientist and less the philosopher. Thus, the more philosophically rewarding (and confusing) discussion of causation fell to Descartes, who shares many of Bacon's reservations about both scholastic inference (the syllogism) and the causal language of Scholasticism. But Descartes, in contrast to Bacon, sometimes seems to want to replace the metaphysical account of Scholasticism with a metaphysical account of his own. In any case, Descartes quickly brings the focus of his discussions of causation to bear on the nature of *efficient* causation. He seldom mentions material causes, and he relegates final causes to the (unknowable) abyss of divine wisdom.

When Descartes turns to the problem of causation he has his eye on three distinct arenas of causal interaction. The first, and the one about which he feels most confident, is the interaction between bodies (Descartes 1985: i. 285). Physics was the science of his time and the laws of motion were much in the mind of Descartes and his contemporaries. But while questions about inertia and what happens in the collision of physical bodies are often taken as the paradigm of a causal interaction, Descartes, the dualist, has equally important questions to ask about how it is that minds are affected by bodies and how minds affect bodies. And, since Descartes is also a theist who believes that God is active in the world, he also seeks to explore God's causal influence in the world. What is it that God does and what is it that is left to finite substances to do?

Throughout his writings Descartes holds steadfastly to a causal principle regarding efficient causation. We will simply call this principle 'Descartes's causal principle'. In one of its simpler versions it is: 'There must be at least as much (reality) in the efficient and total cause as in the effect of that cause' (ibid. ii. 28). Descartes offers several variations on this principle, although all of them concern efficient causation. He suggests that the effect must be 'like' its total cause (Descartes 1992: iii. 340). He claims that there is nothing in the effect that was not previously in the cause either in a 'similar or higher form' (Descartes 1985: ii. 97, 116).

Some scholars have taken Descartes's principle to be a causal *likeness* principle that asserts

that cause and effect must be similar to one another (Watson 1966; Radner 1978; Clatterbaugh 1980). Why must there be a likeness between cause and effect? Historically, the Scholastics had followed Aristotle in holding that the form of anything that comes to be must pre-exist in its cause, and they had toyed with a model of causation called an 'influx model' that explains causation through the transfer of an entity (usually a property) from the cause to the effect (O'Neill 1993). Descartes's own examples of causal interactions lend themselves to such a conception of causation, a conception that Radner (1978) finds throughout much of Descartes's writing. Broadly speaking we can call this view when it is attributed to Descartes 'the transference conception' of causation.

The transference conception is easily characterized. The *realities* of substances are their properties, qualities, modes, accidents, or attributes (Clatterbaugh 1980; Descartes 1985: i. 208–9; ii. 114). The causal principle, as we have noted, states that the realities of the effect must be present in some sense in the total efficient cause. The reason that these realities must be present in the cause is that they cannot come to the effect out of nothing. This kind of transference is explicit in Descartes's rules that govern the dynamics of bodies wherein Descartes (1985: i. 242) frequently speaks of the transference of a quantity of motion from one body to another. Similarly in the *Meditations* Descartes (1985: ii. 28) insists that heat cannot exist in a body except when it is produced by another body with at least at much heat. Heat is the paradigm of a quality that transfers from one body to another and must preexist in the cause. The transference conception becomes even more plausible when one realizes that Descartes is very much a reductionist about physical change. Descartes's reductionism eliminates one argument that someone might offer against the transference interpretation to the effect that obviously Descartes does not require that the efficient cause of a rainbow be itself a multi-coloured thing. For Descartes (1985: i. 83–4, 279) fire, heat, colour, and many other qualities other than size, shape, and motion or rest, are due to the corpuscular motions in bodies. Fire then can blacken wood without being itself black because black is a colour and as such reducible to corpuscles in motion. Thus, if causal interaction among bodies requires a transfer of a quantity of motion from body to body and if all physical change comes about from bodily interaction, then, at least for body–body interaction it looks like the transference conception can be defended as Descartes's conception of causation (Radner 1978: 8). The transference conception also makes it clear why Descartes so often says that the cause must be like its effect, which simply means that the cause, conceived as matter in motion, must possess at least the quantity of motion found in the effect, which makes the cause and effect similar because they share at least this property.

The transference conception as an interpretation of Descartes's view on causation runs into insurmountable difficulties, however, when one turns to body–mind interactions or interactions between God and the world. In talking about sensations, for example, that are caused in the mind by physical bodies, Descartes is clear that there is no resemblance between the sensation in the mind and the physical action of the bodies (1985: i. 165–7). After all, minds do not have the property of motion in a physical sense so they can neither impart nor take on motion from physical bodies. Similarly, God, who is said to have created motion in the world, is not himself in motion. In cases of causal interactions between substances that do not share the same properties, Descartes falls back on a weaker causal principle that says that things can causally interact when they are of the same ontological status or when the cause is

of a higher ontological status than the effect. Such bare ontological relations satisfy the necessary condition for efficient causal interaction. Thus mind and body can interact because they are both created substances even though their natures are totally unlike (1985: ii. 275). And, God can act on the world because God is an infinite substance and hence occupies a higher ontological category (Clatterbaugh 1980).

The transference conception is undermined by two further considerations about motion. The first is that Descartes (1992: iii. 382) himself seems to hold that properties cannot move from one body to another. If he holds such a view, then clearly the transference model is literally impossible (Clatterbaugh 1999: 30–2). Second, and perhaps as damaging is the point defended by Hatfield (1979), Garber (1992), and Clatterbaugh (1999) that Descartes really does not have an ontological place for motion as a force that could be transferred as a property of bodies. Descartes speaks of motion in two ways. Motion is the relocation of a body from one region of space to another over time, but it makes no sense to speak of this translocation as being transferred from one body to another. In the second sense the motion of a body is a force that produces the translocation and may be measured by the quantity of motion in a body. But force is not truly an accident or mode of a body and so is not a candidate to be transferred according to the transference model (Garber 1992; Clatterbaugh 1999: 32).

The failure of the transference conception as a workable model for causation in Descartes has led other scholars to pursue alternative accounts. One that has attracted a fair number of adherents is the view that Descartes really does not believe in the possibility of causal interaction between or among finite substances, or at least some supposed interactions are spurious. On this view God is the sole cause of what look to be interactions between at least some and perhaps all finite substances. Descartes is an occasionalist in whole or part.

The idea that Descartes is an occasionalist goes back at least to the seventeenth century. Malebranche (1980*a*: 466) states that Descartes as well as any who follow the light of reason hold that all motion is but the consequence of the volitions of God. Descartes (1985: i. 202; 1992: iii. 25) himself lends authority to this reading by identifying God as the total and efficient cause of all things. If God is the total and efficient cause of all things, what else is there for finite things to do? But Descartes (1985: i. 93) seems to think that there are things that cannot properly be attributed to God. Descartes (ibid. 285) explicitly attributes motions in some bodies to the actions of other bodies acting upon them. But given what Hatfield (1979) and Garber (1992) say about Descartes's inability to include force or power in his basic ontology, we are left with the conflict that on the one hand Descartes seems clearly to think that bodies act on other bodies and on the other he has no metaphysical story that tells how this could happen.

Scholars who focus their attention on the interaction of mind and body have other reasons for suspecting why Descartes often sounds like an occasionalist (Broughton 1970). In his *Comments on a Certain Broad Sheet* Descartes (1985: i. 304) uses occasionalist language when he says that we judge that things outside the mind transmit signals to the mind that give it the occasion to form ideas by means of the faculty innate to it. The case for an occasionalist reading is even stronger if one believes that Descartes conceives causation as transference. Clearly physical things or even the brain cannot transfer anything, even motion, to the mind, which has no physical presence (Broughton 1970; compare Garber 1992: 22). McCracken (1983: 91–6) notes that in the light of such passages it is easy to find 'the seeds of

occasionalism' in Descartes's accounts of mind–body interaction.

Descartes's *divine concursus* argument in the Third Meditation also points in the direction of an occasionalist reading. According to this argument God continually recreates the universe from moment to moment (1985: ii. 33). If God must recreate the universe and all its contents at each moment then God presumably must create each particular complete with all its properties at each moment. Indeed Descartes (1992: iii. 272) seems to offer just such a view in his correspondence with Elizabeth of Bohemia. At the same time, and in contrast with theses passages, Descartes seems to hold fast to the idea that there are changes that are produced by finite things alone; in fact these changes are inappropriate to God. Just as God points the will towards good, the will can on its own turn the agent towards what is wrong or evil, so God causes rectilinear motion but finite bodies cause other types of motion (1985: i. 92–3, 97, 240, 285).

If Descartes is not an interactionist who embraces a clear conception of causation and he is not an occasionalist although he touches on occasionalist language at certain points, perhaps he is just confused, or more charitably perhaps he is more concerned to identify some of the particularly important causal connections in the world, for example, those that exist between bodies in motion, those that cause sensations and passions and ideas in the mind, and those that enable the will to act on the body and on the world. His goal may be to point out that we have made considerable progress in finding scientific explanations of these many kinds of efficient causation without really being able to say exactly how the causal connection works, metaphysically speaking (Wilson 1991: 133; Sleigh 1990: 172; Bedau 1986: 495). This reading of Descartes would align him with the late moderns who are willing to leave it to successful science to identify the causes of things. Certainly Descartes's commitment to scientific inquiry may mean that he is satisfied with the scientific accounts and therefore avoids any definitive metaphysical account of causation.

3. OCCASIONALISM: MALEBRANCHE AND OTHERS

The 1660s and 1670s saw an upsurge in the number of philosophers who embraced either full or partial occasionalism. Many of these philosophers were perceived as Cartesians or as closely allied to Cartesianism. Geraud de Cordemoy published *Le Discernement de l'ame et du corps* in 1666. In this work he exploits Descartes's impoverished concept of matter. He argues that because the essence of matter is extension, matter could not possibly cause itself to move or to move another (Nadler 1993*b*: editor's introduction, 24–5). Louis La Forge extended his occasionalism at least to the mind–body connections (Nadler 1993*c*). By the early 1670s some began to identify occasionalism with the causal view of Cartesians. A letter in 1672, probably written by the Jesuit A. Rochaon, stated that 'all Cartesians agree that God alone is able to cause motion' (Lennon 1980: 810). Such a perspective on Descartes was surely aided by the fact that both Nicholas Malebranche and Anthony Le Grand regard themselves as Cartesians whose views were faithful to Descartes. Le Grand's English translation of his own work embraces such a point of view in its title, *An Entire Body of Philosophy according to the Principles of the Famous Renate Des Cartes* (1694) (Sorley 1965: 336). Malebranche's various works on occasionalism such as his *Recherche de la vérité* (*Search after Truth*) published in 1674–5 and *Entretiens sur la metaphysique et sur la religion* (*Dialogues on*

Metaphysics and on Religion) published in 1688 embroiled Malebranche in enough controversy that he believed he needed to write his *Elucidations,* which were published with the *Search* and which constitute perhaps the sharpest defence of his views on causation (Malebranche 1980*a*: 530).

   Both Le Grand and Malebranche get to their occasionalist views via arguments that would be easy for a Cartesian to understand. Le Grand (1694: i. 22, 50), for example, clearly holds a version of Descartes's causal principle, namely, 'a cause cannot give that which it hath not'. Le Grand, however, clearly believes that if there were causal interaction between finite substances it would be by transference, but that that is impossible is a view he also attributes to Descartes (ibid. i. 16; iv. 119). He further believes that when Descartes talks of the moving force of bodies he is really talking about the will of God. The causal principle plays only an indirect role in Le Grand's occasionalism. He suggests that if one finite thing were going to influence another it would have to be by transference, which is consistent with his causal principle. However, since accidents are embedded in the very substance that has that accident, that very (individual) accident cannot transfer from one finite substance to another; transference is blocked, it is metaphysically impossible, and therefore causal interaction is impossible (ibid. i. 8; iv. 105). Only God's will gives the appearance of one finite substance acting upon another finite substance.
   Le Grand also finds an argument for occasionalism in Descartes's *divine concursus* argument of the Third Meditation. Here Descartes (1985: ii. 33) argues that God must continually recreate the world anew if it is to continue. If one assumes that in recreating the world anew God must create each substance with all of its features—it is hard to imagine how else *divine concursus* might work—then it is easy to see how one might conclude that God's will and only God's will is the total efficient cause of every single change in the world (compare Le Grand 1694: i. 15; ii. 73). Le Grand (i. 7, 12; ii. 65, 70) further secures this reading of *divine concursus* by identifying God's will with God's knowledge. Thus, if God knows the world in all its changes God also wills the world in all its changes and thereby becomes the cause of everything. And, given the haplessness of bodies (hence they cannot be God's instruments) it seems that the ground has been laid for a thoroughgoing occasionalism, which Le Grand fully embraces.
   Malebranche is more original and less Cartesian than Le Grand, but even so some of his arguments are refinements of the arguments that one can read in Descartes and Le Grand. (Note that although Le Grand published his French occasionalist writings in the early 1670s and Malebranche's *Search* came out a couple of years later, there is little evidence that Malebranche was greatly influenced by Le Grand. Both seem to have read Descartes, however, through the same lens.)
   Malebranche is well aware that for Descartes and most of his contemporaries the paradigm instance of causation is the interaction of bodies in collision. Malebranche (1980*a*: 660) comments on how persuasive it is to see bodies in collision. In those cases his eyes tell him that 'the one is truly the cause of the motion it impresses on the other'. However, Malebranche (ibid. 243–4) is also aware that in Descartes's metaphysics force has no place and that Descartes is unable to explain how matter that has only the properties of size, shape, and translocation can possibly move itself or another bit of matter. This familiar argument,

however, addresses only body–body interaction and Malebranche has a much more ambitious agenda, he would deny all causal interactions except causation by God.

Towards that end Malebranche offers a simple and original argument. A 'true cause' he tells his readers is one 'such that the mind perceives a necessary connection between it and its effect' (ibid. 450). There is only one such cause, namely the will of an all-powerful God. There is some dispute as to whether Malebranche requires logical necessity or metaphysical necessity: it seems quite likely that he means both (Clatterbaugh 1999: 115–17). This necessitarian conception of causation, in one fell swoop, has the effect of making purported instances of causation between or among finite substance illusory. And, since God is all-powerful and it is a corollary of that conception that if God wills something it necessarily happens, Malebranche can conclude that God and only God qualifies as a true cause.

Of course a concurrentist might argue that *both* God and finite substances can be true causes, certainly that claim is central to Aquinas's concurrentism. But Malebranche tries to deflate this view by offering his version of the *divine concursus* argument (ibid. 117–19). Of all those accused of occasionalism or occasionalist sympathies, Malebranche is the most explicit in using this argument to achieve his ends. Simply stated, the argument is that since God must constantly recreate the universe and everything in it in order for things to endure, God must also recreate the properties of each thing along with it. Thus, God, and only God, is causally responsible for each and every state of every substance at every moment (Malebranche 1980*b*: 157; 1980*a*: 678–9). In short, there is nothing left for finite causes to do, or, as Malebranche is fond of saying, 'God needs no instruments to act' (1980*a*: 450).

Just how God implements this occasionalism poses philosophical as well as theological problems. God after all is eternal and changeless. It hardly seems possible that God can be engaged directly with the history of the world, as the *divine concursus* argument seems to suggest. Nor does it seem plausible that God acts only by first creating the laws of nature and then letting nature run itself: Malebranche is no deist (Clatterbaugh 1999: 119–21). Nadler (1993*a*; 1995), Clarke (1995), and Clatterbaugh (1999) all attempt to defend slightly different positions with regard to God's causal involvement with change. Malebranche is clear that God does not violate his own general laws and that God does not have a succession of wills, both would be incompatible with his eternal, unchanging nature (Malebranche 1980*b*: 173, 175). At the same time it is clear especially in the *divine concursus* argument that the contents of divine volitions have particular content. What this suggests is that for Malebranche, God, at the creation, has acts of will whose content is both the general laws of nature *and* the particular states of nature consistent with those laws (Clatterbaugh 1999: 123). In short God does everything and that is done all at once and outside history.

In spite of his belief that God is the only true cause, Malebranche is not afraid to explain changes through what he calls occasional causes or natural causes. His thesis is a metaphysical one and not an effort to revise the language of scientific explanation (Malebranche 1980*a*: 662). But having embraced metaphysical occasionalism Malebranche must give up the idea that a proper scientific explanation is one that presents the real causes of a phenomenon. That separation between real cause and scientific causes proves to be a violation of common sense that Hume later finds unacceptable (Watson 1993).

4. MATERIALISM: HOBBES AND GASSENDI

If Descartes has one foot in occasionalism his other foot is firmly planted in classical mechanism. For all his struggles with his causal principle, the status of force in his metaphysics, and his inability to explain mind–body interaction, in his physics and especially body–body interaction, Descartes is a committed naturalist who holds that body–body interactions are the best understood examples of causal interaction and that the collision of corpuscles whose only properties are size, shape, and translocation is the foundation of most other kinds of physical change.

Descartes's contemporary, Thomas Hobbes, in his mature works, *Leviathan* (1651) and *Concerning Body* (*De Corpore*)(1655), reveals an even more thorough-going materialism. There exist only material substances in space and time—immaterial substance is an oxymoron for Hobbes—and these bodies interact only by contact in accordance with the laws of motion (Hobbes 1839: i. 75–6, 102). Cause and effect for Hobbes are identified with the agent (cause) and the patient (effect), the actor and the acted upon (1839: i. 122; 1948: 71). Hobbes thinks of the *properties* of the agent as the explanatory causes and the *properties* of the patient as the effects and since he seems to hold that unless the cause is present with the effects it cannot be efficacious, causation must be instantaneous (Clatterbaugh 1999: 77).

Even at first glance Hobbes's definition of a cause as the total set of accidents (especially motions) that belong simultaneously to both the agent and patient and are jointly sufficient and individually necessary for the effect is not an especially informative one. The definition itself relies on the concepts of *agent* and *patient,* which makes it circular. Furthermore, this conception of causation does little to enlighten one as to how causation works. The old problem of how to talk about motion, as force or as translocation, reappears in Hobbes's (1839: i. 71, 204) accounts of motion. Hobbes (1994: 7) is clear that 'motion produceth nothing but motion', but again *produce* is itself a causal concept. Such insights offer no explanation of how motion produces motion.

In other places Hobbes (1839: i. 43–4) seems to conflate 'cause' with 'premise' and 'effect' with 'conclusion'. Indeed, Hobbes's desire to reduce everything to matter in motion leads him to identify the entailment relation with a series of corporeal states in the mind wherein the premisses (brain states) are causes that lead one directly to the conclusion or effect (another brain state) (Clatterbaugh 1999: 71–2).

In the end Hobbes displays a thoroughgoing materialism that presents causal interactions among bits of matter as the result of 'stronger' and 'weaker' forces ('endeavours') that compete for 'influence' among the bits of matter (1839: i. 343–5). Hobbes's laws of motion are largely geometrical theorems and his stories about how macro-phenomena reduce to corpuscular motions and how those small bits of matter interact is largely metaphorical. But causation itself was for Hobbes reduced to matter in motion, however incomplete his account.

Petrus Gassendi was a friend and correspondent of Hobbes and one of the critics of the *Meditations.* Gassendi's views are represented in the Fifth Set of Objections. His *Syntagma Philosophicum,* a philosophical account of his atomistic physics, was completed in 1655 just before his death and published in 1658. Like Hobbes, Gassendi rejects final causes and formal causes, although he wavers a bit on the value of final causes only because he believes that the universe reveals the power and goodness of God (Descartes 1985: ii. 215; Gassendi 1972: 208–9). Formal causes are just incorporeal efficient causes whose principle of action is occult and

'mere verbiage' (ibid. 415–16). Gassendi directly challenges Descartes's causal principle, especially in the form that suggest a likeness between the cause and the effect. His point is that although in some cases the efficient cause resembles its effect, this similarity is by no means a necessary condition as for example when an artefact is produced by the skill of an artisan (Descartes 1985: ii. 201). Gassendi (1972: 410–11) relegates material causes to the patient being acted upon by the efficient cause. So, as with Hobbes and to a large extent Descartes, the only cause that really deserves mention in science is the efficient cause conceived in Gassendi's philosophy as impenetrable atoms in motion (ibid. 414).

Like Hobbes and Descartes, Gassendi is uncertain what to do with motion as force or power. It is not an intrinsic property of atoms the way that size, shape, translocation, and solidity are (Gassendi 1972: 424). Motion, as force or power, is something that is imparted to atoms at the time of creation and preserved by God (ibid. 99). God's role for Gassendi is to preserve atoms and to import motion to them, at least at the beginning. Gassendi's materialistic view of causal interaction leads him to deny mind–body interactions (Descartes 1985: ii. 239; Gassendi 1972: 414) and to focus on body–body collisions. Thus, Gassendi, apart from his commitment to atoms as opposed to infinitely divisible corpuscles, offers a picture that is very similar to Descartes's. Change in the material world is due to solid bodies in collision, force is something alien to bodies, something that is imparted to them by God, and mind–body interaction is mysterious and unintelligible. Causation, with a brief nod to God's design, is exclusively efficient materialistic causation.

Gassendi is more of a sceptic than Descartes. He does not embrace innate ideas and at one point suggests that we should be satisfied if we are able to glimpse not truth but some 'slight image' of it or even 'its shadow' (Gassendi 1972: 327). And, although he has confidence in reason and sees reason as a corrective for inadequate sensory knowledge, Gassendi still seems to embrace the Baconian idea that the individual and particular is better known than the universal and that induction, which proceeds from the particular, is the appropriate method in science, that is, a way to gain insight into things hidden (ibid. 332–6). In particular Gassendi (ibid. 339) embraces a theory that posits atoms, pores, and humours because in the end they do a pretty good job of explaining the appearances of things. But Gassendi's conviction, like Hobbes's geometrical materialism, does little to solve the metaphysical puzzles about how causation actually works in the world to connect certain things and not others. Yet, like Hobbes, Gassendi brings to the fore the idea that particular causal interactions among bodies are the paradigms of causal interactions and that these may well be the best (or only) understood cases of causal interaction.

5. SUFFICIENT REASON: SPINOZA AND LEIBNIZ

The idea that there are no causal relations among finite substances and that all real explanations are in terms of divine will obviously fails to satisfy the demands of a science that increasingly looks for detailed causal accounts of such things as why bodies move as they do when they collide, why and how heat is transferred from one body to another, or why the colours of the rainbow have the order that they do. At the same time the materialist account of matter in motion, while promising, obviously requires that either one become a materialist or that many apparent causal interactions between mind and body are totally unintelligible. Both

accounts fail to measure up to the rationalist hope that there exists a reason why *everything* is the way it is and not some other way. Both Benedict Spinoza and Gottfried Leibniz try to shore up this side of the causation debate with systems that are not only deterministic—everything is caused with the possible exception of God—but also everything is necessitated, that is, causes necessarily bring about their effects.

As Bennett (1996: 61) notes, both Spinoza and Leibniz are committed to 'explanatory rationalism', that is, they believe that there is a satisfactory answer to every 'Why?' question.

Spinoza often attacks the notion of final causes, which continue to have a vestigial role in Descartes and Gassendi. For Spinoza (1985: 422) final causes, either intrinsic or extrinsic, are nothing but 'human fictions'. Descartes's tendency to identify final causes with divine purposes—a view that Leibniz also favours—fails in Spinoza's view for the simple reason that God has no purposes, hidden or otherwise (ibid. 439). Only efficient causation serves to explain both the existence and the essence of things (ibid. 431). For many Spinoza commentators (Parkinson 1954; Curley 1969; Allison 1975; Bennett 1996) Spinoza's necessitarianism quickly follows from his views on causation. Simply stated, Spinoza's view is that God or Nature exists necessarily and that all things follow from God or Nature in the same way as the theorems about triangles follow from the nature of triangle (Spinoza 1985: 426). Thus, on this reading of Spinoza, there is no distinction between ground and consequent and cause and effect. We have already noted a similar slide in Hobbes who tries to ground entailment in natural cause-and-effect relations, but while Hobbes pushes to the side of material causation and thereby loses entailment, Spinoza in some passages seems to go the other way, he loses natural causation to entailment.

But much of Spinoza's writing in the *Ethics* seems to run counter to this inferential view of causation, and furthermore, there may be more than one way to interpret the inferential reading of Spinoza (Clatterbaugh 1999: 135–9). One of the first problems with the inferential view is that it does not seem to make room for Spinoza's distinction between immanent and transitive causes of all things (Spinoza 1985: 428; Wilson 1991). And Spinoza (1985: 427) is clear that particular things (finite modes) are transitive causes of other particular things while God is the immanent cause especially of timeless things such as essences (Clatterbaugh 1999: 467). And, epistemically, God is required in every explanation since everything is conceived through God. But a full causal account of any finite mode must make reference to other finite modes (transitive causes), indeed to an infinite number of finite modes (Spinoza 1985: 428). Thus, it is hard to see how God, as immanent cause, can be the *full* explanation of any finite mode.

Even those like Curley (1969: 73) who take Spinoza to be an explanatory rationalist take God to be only a partial cause of finite things. This is supported by the fact that Spinoza says in Definition I of Part II of the *Ethics* that an adequate cause is one through which its effect can be clearly and distinctly perceived (1985: 492). But starting with God one only gets clear and distinct perceptions of general truths and not of particular things or finite modes (ibid. 467).

When it comes to interactions between and among bodies Spinoza is as much a mechanist and materialist as are Descartes, Hobbes, and Gassendi. His conception of body is that each body is divided into smaller bodies and all changes of state come about through the motion or

rest of the parts. All of Nature can be conceived as a body composed of bodies, he says in Lemma VII of Part II (ibid. 462). As to the interaction of bodies with minds Spinoza seems to opt for a dual aspect theory in which mind and body are regarded as two different ways of viewing and talking about one and the same thing, at least this is what he seems to say in Part III, Proposition 2 (ibid. 494).

The apparent breakdown of the inference view of causation might suggest that Spinoza is not a necessitarian or could escape necessitarianism. However, it is not clear that Spinoza wants to avoid this consequence, indeed he seems to embrace the principle of sufficient reason that there is a reason for everything being the way it is and that that reason necessarily determines its consequent (ibid. 417). There are many ways to see Spinoza's necessitarianism but the easiest is to take Spinoza's word in the Preface to Part IV that God and Nature are one and that God or Nature is a necessary being, from which it seems to follow that all of Nature is necessary as well (ibid. 544). Contingency for Spinoza is an epistemological category, something is contingent only to the extent that one is ignorant of its causes (Delahunty 1985: 161).

Leibniz, too, is a necessitarian, and he too embraces explanatory rationalism. He holds that there is a sufficient reason for everything and that that reason is itself a necessary truth (Leibniz 1989: 33, 47; 1969: 337, 464). But the ground of Leibniz's necessitarianism is very different from Spinoza's. Leibniz's ground is not a necessary God or Nature, but a concept containment theory of truth (Leibniz 1969: 337; Mates 1986: 153–4; Clatterbaugh 1973). Simply stated, for every true proposition the subject concept contains the predicate concept, which commits Leibniz to the view that every true proposition is necessarily true (Leibniz 1969: 337, 646). While Leibniz's view entails that there exists a rational explanation for every truth, it does not entail that finite minds can have such an explanation. In fact since truths about individual things require an infinite analysis that precludes finite minds from having ever having a full explanation, only God has an answer to every 'Why?' question (Adams 1994; Sleigh 1990: 170).

In spite of the common ground between Leibniz and Spinoza there are important differences concerning causation. Leibniz (1989: 55) thinks that final causes are useful; knowing final causes—these being God's guiding principles of plentitude, perfection, and simplicity—may provide a short cut to discoveries about the world. Also final causes are useful as a way to explain why the world has the lower-level natural laws that it does (ibid. 211).

When it comes to interactions between and among bodies or between minds and bodies, Leibniz, like Malebranche, is willing to talk as if there is interaction, but strictly and metaphysically speaking there is no causal interaction between bodies, between minds, or between minds and bodies. Each substance has an internal force or nature that causes it to play out its various states in harmony with all the other substances in creation (ibid. 144). The world so organized by a pre-established harmony looks no different from a causally interactive world, it is just that interaction is not what is going on (Sleigh 1990: 162). Thus, although we may talk about the efficient cause of any particular phenomenon or change in a substance as if it were caused by another substance, that is simply a manner of speaking in conformity with ordinary language; what is going on metaphysically is that each change in any substance is a consequence of the internal state of each substance (Leibniz 1969: 495; 1989: 81–2). Each individual thing is an automaton driven to change only by its own internal force

(Bobro and Clatterbaugh 1996). And that sequence of states driven by the internal force is uniquely determined or explained in turn by the complete individual notion of that substance (Leibniz 1989: 47). Leibniz takes as much pride in his *New System of Nature* and this theory of pre-established harmony as Malebranche did in his occasionalism. But both pay a price. They are required to give lip service to the idea that things causally interact while denying that that is the real story.

6. AGAINST METAPHYSICS: BOYLE, NEWTON, AND LOCKE

By the late seventeenth century the new science was posting considerable success with respect to both its claims of scientific discovery and its impact on the philosophical discussion of causation. Even the avowed Cartesian Jacques Rohault, whose *A System of Natural Philosophy* was published in Paris in 1671, ignores the metaphysical worries about the nature of the motive force and the role of God in explanation (Garber 1992: 62). In his author's preface Rohault (1723) warns that natural philosophy is held back by treating of matters 'too metaphysical' and disputes about abstract questions that are of no use in explaining particular effects in nature. Matters metaphysical include concerns about the nature of the motive force, the existence of the vacuum, or the existence of subtle matter.

Robert Boyle arrives at many of the same conclusions in his own scientific writings. Boyle's essays attempt to defend a view of scientific explanation that avoids final causes, accidental forms, occult properties, and above all metaphysical speculation. Boyle tries to be as metaphysically neutral as possible. His scientific explanations are naturalistic but he avoids the controversy about atoms versus corpuscles, he avoids positing powers, he dismisses final causes or divine design explanations of particular changes, and he tries to dodge questions about the vacuum and subtle matter (Boyle 1979: 38–9, 45). Boyle can be read as a 'restorer' of the mechanical philosophy of Hobbes and Gassendi with a wary eye on the metaphysics that have entrapped many of the scientist/philosophers of his and earlier generations (Boas 1952; Joy 1987). Boyle tries to restrict his explanations to the empirically available properties of size, shape, solidity, and motion or rest. It is worth comparing Boyle and Rohault on the merits and demerits of the Chemists' analysis of matter (the Chemists were a school that identified sulfur, salt, and mercury as the building blocks of other substances; Clatterbaugh 1999: 169–70). Both find the view too superficial, not deep or radical enough, too metaphysical in its initial hypotheses, and incompatible with observations readily made in the laboratory.

Except for his struggle with the force of gravity, Isaac Newton clearly belongs in the same grouping as Rohault and Boyle. He aspires to an account of nature that follows the mechanical principles (Newton 1953: 10–11). But Newton's science was much more sophisticated than his mechanistic predecessors. Unlike Gassendi he could not treat the attractive forces as a series of hooks and cables that entwined matter. Yet, Newton (1953: 125–6) bristles at the attempts to label gravity an occult power; he notes that it is occult in the sense that it can not be seen and is not fully understood, but, he argues, that it is not occult in the sense that the laws of gravity are not understood. Gravity is one of the 'active principles' of nature whose existence is justified by the precision with which it allows one to explain the behaviour of particular

phenomena (ibid. 54, 176). But while gravity may be an unexplained explainer, Newton (ibid. 5) does not seem to doubt that in the end it too will have a simple explanation and for now it is the best hypothesis available.

It is commonplace in Locke scholarship to treat John Locke as a philosopher guided by Boyle (Alexander 1985; Ayers 1991; Keating 1993). Locke in *An Essay Concerning Human Understanding* (1689) makes it clear that he seeks to avoid speculations about the metaphysical structure of the world. He prefers to see himself as an 'Under-Labourer' who seeks to remove some of the rubbish that lies in the way of knowledge (Locke 1975: 9–10). Locke is content, in most of the *Essay*, to restrict himself to an introspective examination of the ideas that he finds in the mind; his concern is what ideas are there, how to break down the complex ones, and how they serve to give us knowledge, where knowledge is described as the perception of the agreement and disagreement of ideas (ibid. 525).

What sometimes makes Locke maddeningly complicated is that he does not stick to his purported strategy. Instead he wanders off into speculations about the causes of these ideas (qualities), the forces in bodies, and the existence of substance in general and the existence of real essences in things. Worse yet, he sometimes suggests that the true causes of things are hidden from us (McCracken 1983: 150). And standing Boyle on his head Locke treats power as an undefined, simple idea in terms of which he defines quality (the power to produce ideas in us) (Clatterbaugh 1999: 186–91). Thus, scholars are left to debate what Locke means by non-ideas such as substance in general and real essence, all the while wondering what these have to do with knowledge, which is described as the perception of the agreement and disagreement of ideas and nothing more.

Locke's contribution to the modern discussion of causation is modest and not particularly helpful. In the *Essay* Book II, chs 21 and 26 he discusses the notion of power and the concepts of cause and effect. But there is no real metaphysical probing as to what is going on in causation and the exercise of power. He suggests that we get our idea of power from the will (1975: 235) as well as from watching things change. But, perhaps in the anti-metaphysical spirit of his time, Locke restricts himself to giving some examples of causal interaction and saying simply (and vacuously) that the cause is what produces something and the effect is what is produced.

Locke's reluctance to engage in metaphysical speculation and his willingness to take as causes what contemporary scientists posit as causes is indicative of the return to the Baconian view. By the time of Locke the multitude of causes inherited by the early moderns has nearly evaporated. The dogmas of philosophers that purportedly held back science or at least confused scientific explanation have been abandoned. Locke, Boyle, Rohault, and Newton share this perspective. While they are not quite as militantly anti-metaphysical as the logical positivists of the 1940s, they do believe in an access to nature that is relatively free of the metaphysical speculations of Descartes, Spinoza, Malebranche, and Leibniz. They are prepared to accept as causes what scientists identify in successful scientific explanations. End of story. The early moderns prepare the ground for Hume's observation that what Malebranche called the occasions or natural causes of things *are* the causes of these things and not something else. They also prepare the ground for Kant's worry that the concept of necessity needs to be put back into the concept of causation and that philosophy should not be left out of the discussion about causation.

FURTHER READING

McCracken's study of Malebranche's influence on British Philosophy (1983) and Watson's book on the downfall of Cartesianism (1966) provide a good beginning to the study of causation in early modern philosophy. Loeb (1981) and Clatterbaugh (1999) both try to provide an overview or big picture of the controversies and dominant themes of the period. The collections of papers in Cover and Kulstad (1990) and Nadler (1993*b*) offer more detailed scholarly accounts of particular philosophers of the period.

REFERENCES

ADAMS, R. M. (1994). *Leibniz: Determinist, Theist, Idealist*. Oxford: Oxford University Press.

ALEXANDER, P. (1985). *Ideas, Qualities, and Corpuscles: Locke and Boyle on the External World*. Cambridge/New York: Cambridge University Press.

ALLISON, H. E. (1975). *Benedict de Spinoza*. Boston: Twayne.

AQUINAS, T (1945). *The Basic Writing of Saint Thomas Aquinas*, trans. and ed. A. C. Pegis. New York: Random House.

AYERS, M. (1991). *Locke*. 2 vols. London: Routledge.

BACON, R. (1889). *Novam Organum*, ed. T. Fowler. Oxford: Oxford University Press.

BEDAU, M. (1986). 'Cartesian Interaction', *Midwest Studies in Philosophy* 10: 48143–502.

BENNETT, J. (1996). 'Spinoza's Metaphysics', in D. Garrett (ed.), *The Cambridge Companion to Spinoza*. New York: Cambridge University Press, 61–88.

BOAS, M. (1952). 'The Establishment of the Mechanical Philosophy', *Osiris* 10: 412–541.

BOBRO, M., and CLATTERBAUGH, K. (1996). 'Unpacking the Monad: Leibniz's Theory of Causality', *The Monist* 79/3: 408–25.

BOYLE, R. (1979). *Selected Philosophical Papers of Robert Boyle*, ed. M. A. Stewart. Manchester: Manchester University Press; New York: Barnes & Noble.

BROUGHTON, J. (1970). 'Reinterpreting Descartes on the Notion of Union of Mind and Body', *Journal of History of Philosophy* 16: 23–32.

—— (1986). 'Adequate Causes and Natural Change in Descartes's Philosophy', in Alan Donagan, Anthony Perovich, and Michael Wedin (eds.), *Human Nature and Natural Knowledge*. Dordrecht: D. Reidel, 107–27.

CLARKE, D. M. (1995). 'Malebranche and Occasionalism: A Reply to Steven Nadler', *Journal of the History of Philosophy* 33/3: 499–504.

CLATTERBAUGH, K. (1973). *Leibniz's Doctrine of Individual Accidents*. Studia Leibnitiana 4. Weisbaden: Franz Steiner.

—— (1980). 'Descartes' Causal Likeness Principle', *Philosophical Review* 89/3: 379–402.

—— (1999). *The Causation Debate in Modern Philosophy* 1637–1739. New York: Routledge.

COLLINS, J. (1960). *God in Modern Philosophy*. London: Routledge & Kegan Paul.

COVER, J. A., and KULSTAD, M. (1990). *Central Themes in Early Modern Philosophy*. Indianapolis: Hackett.

CURLEY, E. (1969). *Spinoza's Metaphysics: An Essay in Interpretation*. Cambridge, Mass.: Harvard University Press.

DELAHUNTY, R. J. (1985). *Spinoza*. London: Routledge & Kegan Paul.

DESCARTES, R. (1985). *The Philosophical Writings of Descartes*, i and ii, trans. and ed. J. Cottingham, R. Stoothoff, and D. Murdoch. Cambridge; Cambridge University Press.

—— (1992). *The Philosophical Writings of Descartes*, iii, trans. and ed. J. Cottingham, R. Stoothoff, D. Murdoch, and A. Kenny. Cambridge: Cambridge University Press.

GARBER, D. (1992). *Descartes' Metaphysical Physics*. Chicago: University of Chicago Press.

GASSENDI, P. (1972). *The Selected Works of Pierre Gassendi*, trans. and ed. C. B. Brush. New York: Johnson Reprint.

HATFIELD, G. C. (1979). 'Force (God) in Descartes' Physic.', *Studies in History and Philosophy of Science* 10/2: 113–40.

HOBBES, T. (1839). *The English Works of Thomas Hobbes of Malmesbury*, ed. W. Molesworth. 11 vols. London: Bohn.

—— (1948). *The Metaphysical System of Hobbes in Twelve Chapters from Elements of Philosophy concerning Body Together with Briefer Extracts from Human Nature and Leviathan*, ed. M. W. Calkins. La Salle: Open Court.

—— (1994) *Thomas Hobbes: Leviathan*, ed. E. M. Curley. Cambridge: Hackett.

JOY, L. S. (1987). *Gassendi the Atomist: Advocate of History in an Age of Science*. Cambridge: Cambridge University Press.

KEATING, L. (1993). 'Un-Locke-ing Boyle: Boyle on Primary and Secondary Qualities'. *History of Philosophy Quarterly* 10/4: 305–23.

LE GRAND, A. (1694). *An Entire Body of Philosophy, according to the Principles of the Famous Renate de Cartes in Three Books: I The Institution, II. The History of Nature, III, The dissertation of the Want of Sense and Knowledge*. London: Samuel Roycroft.

LEIBNIZ, G. W. (1969). *Philosophical Papers and Letters*, trans. and ed. L. L. Loemker. 2nd edn. Dordtecht: Reidel.

—— (1989). *G. W. Leibniz Philosophical Essays*, trans. and ed. R. Ariew and D. Garber. Indianapolis: Hackett.

LENNON, T. M. (1980). 'Philosophical Commentary', in Malebranche 1980*a*: 757–848.

LOCKE, J. (1975). John Locke: *An Essay Concerning Human Understanding*, ed. P. H. Nidditch. Oxford: Clarendon.

LOEB, L. E. (1981). *From Descartes to Hume: Continental Metaphysics and the Development of Modern Philosophy*. Ithaca: Cornell University Press.

MCCRACKEN, C. (1983). *Malebranche and British Philosophy*. Oxford: Oxford University Press.

MALEBRANCHE, N. (1980*a*). *Malebranche: The Search after Truth/Elucidation of the Search after Truth*, trans. and ed. T. M. Lennon and P. J. Olscamp. Columbus: Ohio State University Press.

—— (1980*b*). *Nicolas Malebranche: Entretiens sur la Metaphysique (Dialogues on Metaphysics)*, trans. and ed. W. Doney. New York: Abaris.

MATES, B. (1986). *The Philosophy of Leibniz*. Oxford: Oxford University Press.

NADLER, S. (1993*a*). 'Occasionalism and General Will in Malebranche'. *Journal of the History of Philosophy* 31/1: 31–47.

—— (ed.) (1993*b*) . *Causation in Early Modern Philosophy*. University Park: Pennsylvania State University Press.

—— (1993*c*). 'The Occasionalism of Louis de la Forge', in Nadler 1993*b*: 57–73.

—— (1995). 'Malebranche's Occasionalism: A Reply to Clarke', *Journal of History of Philosophy* 33: 505–8.

NEWTON, I. (1953). *Newton's Philosophy of Nature: Selections from his Writings* , ed. H. S. Thayer. New York: Hafner.

O'NEILL, E. (1993). 'Influxus Physicus', in Nadler 1993*b*: 27–55.

PARKINSON, G. H. R. (1954). *Spinoza's Theory of Knowledge*. Oxford Clarendon.

RADNER, D. (1978). *Malebranche*. Amsterdam: Van Gorcum.

ROHAULT, J. (1723). *Rohault's System of Natural Philosophy, with S. Clarke's Notes* , ed. and trans. J. Clarke and S. Clarke. New York: Johnson Reprint.

SLEIGH, R. C., Jr. (1990). 'Leibniz on Malebranche on Causality', in Cover and Kulstad 1990: 161–93.

SORLEY, W. R. (1965). *A History of British Philosophy to* 1900. Cambridge: Cambridge University Press.

SPINOZA, B. (1985). *The Collected Works of Spinoza*, i, trans. and ed. E. M. Curley. Princeton: Princeton University Press.

WATSON, R. A. (1966). *The Downfall of Cartesianism,* 1673–1712. The Hague: Martinus Nijhoff.

—— (1993). 'Malebranche, Models, and Causation', in Nadler 1993*b*: 75–91.

WILSON, M. (1976). *Descartes*. London: Routledge & Kegan Paul.

—— (1991). 'Spinoza's Causal Axiom (Ethics I, Axiom 4)', in Y. Yovel (ed.) (1991), *God and Nature in Spinoza's Metaphysics*. Leiden: Brill, 133–60.

CHAPTER 4
HUME

DON GARRETT

All kinds of reasoning consist in nothing but a comparison, and a discovery of those relations, either constant or inconstant, which two or more objects bear to each other.

A Treatise of Human Nature 1.3.2.2; SBN 73[1]

1. INTRODUCTION

No one is more famous for having a theory of causation than David Hume. But what exactly is that theory? Opinions vary. Much of the recent debate about Hume's theory of causation has been provoked by striking readings of Hume as a causal realist—readings that exemplify what is now (following Winkler (1991)) commonly known as 'The New Hume' interpretation. (For a summary description of the debate and a collection of important contributions to it, see Read and Richman 2000.)

Some have taken him to be a *projectivist*. For he holds that a 'necessary connexion' is required as an 'essential ... part' of the relation of cause and effect (THN 1. 3. 6. 3; SBN 87) and then goes on to argue that what we take to be such a 'necessity and power ... in the objects' is in fact merely an internal feeling of 'the determination of the mind, to pass from the idea of an object to that of its usual attendant' (THN 1. 3. 14. 25; SBN 167). The mind, he holds, erroneously treats this feeling as a quality of the objects observed (THN 1. 3. 14. 25; SBN 167), even though the ideas derived from this feeling of determination 'represent not any thing, that does or can belong to the objects' (THN 1. 3. 14. 19; SBN 164). Thus, it seems that he regards ascriptions of causal relations as projections or expressions of internal feelings or attitudes, so that genuine causal relations find no place in the basic metaphysical explanatory structure of the universe.

Others have taken him to be a *reductionist*. For he offers what he calls a 'precise definition' of *cause* as 'an object precedent and contiguous to another, and where all objects resembling the former are placed in like relations of precedency and contiguity to those objects, that resemble the latter' (THN 1. 3. 14. 30–1; SBN 169–70). Moreover, he regularly cites and employs this definition, claiming that its satisfaction is sufficient for causal necessity (THN 1. 3. 14. 33; SBN 171), including that of human actions (THN 2. 3. 1. 1–18; SBN 399–407); and he insists that when we try to speak instead of the necessary connection of causes as something residing in the objects themselves, we either 'contradict ourselves, or talk without a meaning' (THN 1. 4. 7. 5; SBN 267; see also THN 1. 3. 14. 14; SBN 162; and THN 1. 3. 14.

27; SBN 168). Thus, it appears that he employs a method of semantic analysis to conclude that the causal relation is nothing more than 'constant conjunction'.

Still others have taken him to be a *realist*. For he goes on to allow that the same definition may 'be esteem'd defective, because drawn from objects foreign to the cause' (THN 1. 3. 14. 31; SBN 170), and he alludes variously to 'the power by which one object produces another' (THN 1. 3. 1. 1; SBN 69), the 'internal structure or operating principle of objects' (THN 1. 3. 14. 29; SBN 169), and 'the ultimate connexion of ... objects'. Moreover, he concludes that 'we can never penetrate so far into the essence and construction of bodies, as to perceive the principle, on which their mutual influence depends' (THN 2. 3. 1. 4; SBN 400) and that 'we cannot penetrate into the reason of the conjunction' of causes and effects (THN 1. 3. 6. 15; SBN 93). Thus, he seems to affirm that we can and do refer to perfectly real—even if perhaps epistemically inaccessible—causal powers and relations that go beyond both the projection of internal sentiments or attitudes and mere constant conjunction.

While I have quoted only from the *Treatise*, similar passages supporting each of these three interpretations occur in *An Enquiry concerning Human Understanding*[2] as well—sometimes with even greater frequency. The apparent dialectical stand-off among these interpretations has led some to conclude that Hume has no consistent theory of causation at all, but only a confused and contradictory collection of remarks. Still, one of Hume's prime methodological principles is this: where philosophical confusion reigns, cognitive psychology may help to bring resolution. In what follows, therefore, I will try to approach the question of his theory of causation through two questions of Humean cognitive psychology. First, what is his theory of causal inference, particularly as it concerns necessity? Second, what is his theory of causal judgement, particularly as it concerns the deployment of concepts? The answers to these questions, I will argue, reveal that Hume has a reasonably sophisticated 'causal sense' theory of causal psychology that allows him to concede something to each of projectivism, reductionism, and realism without falling into a simple version of any of these epistemological/semantic/metaphysical packages.

2. HUMEAN CAUSAL INFERENCE

In order to understand the relation of cause and effect, Hume asserts, we must first understand the nature of causal inference. Indeed, causal inference *is* the 'discovery' of causal relations, rather than simply a consequence of it, in his view (THN 1. 3. 2. 2; SBN 73); and it is because 'the nature of the relation depends so much on that of the inference' that he is obliged 'to advance in this seemingly preposterous manner' of examining causal inference before explaining the nature of the causal relation (THN 1. 3. 14. 30; SBN 169).

The main outline of the theory of causal inference that Hume presents is well known. All causal inferences depend, in one way or another, on experience of constant conjunctions. Following experience in which objects or events of one kind are generally or always followed by objects or events of another kind, the mind develops, through 'custom or habit', a propensity to form an 'idea' of an object or event of one of these kinds upon having a 'perception' (which is itself either an 'impression' or an idea) of an object or event of the other. Where the newly occurring perception has sufficient force and vivacity—typically,

when it is an impression or an idea of memory—the associated idea formed by the mind acquires some of this force and vivacity, thereby constituting it as a *belief*.

But this is not all. In the course of making this transition, Hume asserts, the mind perceives a distinctive internal impression—that is, an 'impression of reflection' rather than an 'impression of sensation'. He characterizes this impression as being 'of determination', since one in fact feels the impression whenever the mind makes, or is about to make, an involuntary custom-or-habit-based inference. (Of course, the regular conjunction of this feeling with the corresponding inference, like all constant conjunctions, is itself learned only by experience.) He also, however, characterizes it as an impression 'of necessary connexion', for he holds that the application of the term 'necessity' indicates the mind's determination to conceive things in a certain way—that is, its inability to conceive and affirm otherwise. This is so whether the necessity in question is that of what he calls 'relations of ideas' (which are always either intuitively self-evident or demonstrable) or that of causes. Thus he writes:

As the necessity, which makes two times two equal to four, or three angles of a triangle equal to two right ones, lies only in the act of the understanding, by which we consider and compare these ideas; in like manner the necessity or power, which unites causes and effects, lies in the determination of the mind to pass from the one to the other. (THN 1. 3. 14. 23; SBN 166)

In the latter case, he observes, 'the objects seem so inseparable, that we interpose not a moment's delay in passing from the one to the other' (THN 1. 3. 8. 13; SBN 104).

Yet although necessary relations-of-ideas and necessary causal relations both involve an inability to think otherwise, there remains a crucial difference between the two kinds of necessity. In the case of relations-of-ideas, which include purely mathematical truths, the inability to conceive otherwise is grounded in the intrinsic character of the ideas themselves, which—at least when they are adequate representations (see THN 1. 2. 2. 1; SBN 29)—represent the intrinsic characters of their objects. In the case of causal necessity, in contrast, the inability is more properly a psychological difficulty in separating two perceptions, and a psychological inability to believe their objects to be separated, resulting not from the intrinsic characters of the objects as one conceives them but from the habitual association between them that has been established by constant conjunction. Confusion results when one conflates these two different species of necessity. Thus Hume writes:

'Tis natural for men, in their common and careless way of thinking, to imagine they perceive a connexion betwixt such objects as they have constantly found united together; and because custom has render'd it difficult to separate the ideas, they are apt to fancy such a separation to be *in itself* impossible and absurd. (THN 1. 4. 3. 9; SBN 223; emphasis added)

Much of the source of this careless mistake lies in the mind's natural propensity 'to spread itself on external objects, and to conjoin with them any internal impressions which they occasion, and which always make their appearance at the same time that these objects discover themselves to the senses' (THN 1. 3. 14. 25; SBN 167). Just as the mind tends to mislocate its

non-spatial impressions of sounds and smells in the space of their sources, so too it often mislocates its impressions of necessary connection in the causes and effects themselves, leading to the mistaken supposition that causal necessity is a directly observable feature of those objects. Nevertheless, it remains true that, much as one discovers the *mathematical relations-of-ideas necessity* of twice two making four when one is determined by the intrinsic character of the ideas to conceive of four in the act of conceiving twice two and finds oneself unable to conceive twice two making any other quantity, so one discovers the distinctive *necessity of causes* by being psychologically determined, following the observation of genuine constant conjunction, to infer the existence of one object from that of another and by finding it difficult even to think of the one object without thinking of the other. It is precisely in making such necessity-involving associations and inferences that one first comes to represent two objects as being causally related.

3. Humean Causal Judgement

The basic process of causal inference occurs in many animals just as it does in human beings, Hume emphasizes, and they all thereby 'discover' the causal relations between objects. Only human beings, however, go on to make explicitly conceptualized judgements of the form 'Object *A* and Object *B* are cause and effect.' In order to understand how such explicitly conceptualized causal judgements are possible for Hume, we must understand his view of how explicitly conceptualized judgements concerning qualities and relations are possible at all.

For Hume, perceptions in the mind—ideas, and also many impressions—can represent their objects as having many different qualities. Often, they do so by having those very qualities themselves. For example, perceptions may represent their objects as being round and a particular shade of red, often by being round and that particular shade of red themselves. Similarly, perceptions may represent their objects as standing in many different relations. For example, a pair of perceptions may represent one object as similar in shape but double in size to another—often, once again, by standing in these relations themselves. In his view, mental representation of this kind can occur without explicit concepts. For he recognizes a most basic form of belief that does not require any acts of predication involving explicit concepts but is instead simply an imagistic mental model that possesses sufficient psychological force and vivacity to affect voluntary behaviour—what he also calls a 'lively idea' (see especially THN 1. 3. 7. 5 n.; SBN 96–7). Thus, an animal or human may use a mental model of its environment in order to track and anticipate the movements of its prey; and in virtue of doing so, it may have unconceptualized beliefs about them. While the use of such a model certainly involves capacities or propensities to respond differentially to differences in features of the model— basic capacities or propensities that some might characterize as 'implicit concepts'—it requires no explicit concepts at all. In order to think with generality *about* qualities and relations—say, about all shades of red—or about the classes of things that have them, however, it *is* necessary, on Hume's view, to form explicit concepts of them. As it happens, he has a developed theory of explicit concepts; they are what he calls 'abstract ideas'.

Hume begins *Treatise* 1. 1. 7, 'Of abstract ideas', by arguing that no idea is general or indeterminate in its own nature. In order to have thoughts with generality, therefore, the mind

must employ what he calls the 'imperfect' device of 'abstract ideas' (THN 1. 1. 7. 1; SBN 36). Such ideas arise in the mind, on his view, in the following way. When perceptions resemble one another in some respect, it is natural to apply the same term to each, notwithstanding their difference. Later uses of the term come to elicit a particular and determinate idea—which we may call the 'exemplar'—together with a disposition to revive and 'survey', as needed for reasoning or other purposes, the other ideas whose objects resemble one another in the operative respect (THN 1. 1. 7. 7–8; SBN 20–1). We may call these other ideas, together with the exemplar, the 'revival set'. Thus, the abstract idea of the quality red is a determinate idea of a particular thing having a particular shade of red but associated with the word 'red' in such a way that the mind is disposed to revive and survey any of a set of other ideas of red things for use as needed. If, for example, one's abstract idea exemplar of red is an idea of a scarlet circle, and the claim is made that all red things are scarlet circles, the ideas of red things of other shades or other shapes will (if all goes well) immediately come to mind, allowing one to reject the claim proposed. Similarly, we may suppose, the abstract idea of a *relation* consists, for Hume, of a determinate idea of an ordered pair or other group of particular things taken to stand in that relation and associated with a word in such a way that the mind is disposed to revive and survey for use as needed any of a set of ideas of other pairs or groups whose objects are taken to be similarly related. That this *is* the theory that he applies to the specific concept of the causal relation seems to be confirmed by his remark that 'we must not here be content with saying, that the idea of cause and effect arises from objects constantly united; but must affirm, that *it is the very same with the idea of these objects*' (THN 2. 3. 1. 16; SBN 405; emphasis added).

All belief, for Hume, is a matter of the force and vivacity (that is, liveliness) of ideas that allows them to affect action. Hence, to make an explicit conceptual judgement that a particular object has a given conceptualized quality—and thereby to predicate the quality of the object explicitly—must be to include a lively idea of that object within the revival set of the abstract idea of that quality. To judge explicitly that two particular objects stand in a specified relation must likewise be to include a lively idea of that pair of objects in the revival set of the abstract idea of that relation. To judge explicitly that two objects *A* and *B* stand in the specific relation of cause and effect, therefore, must be to include a lively idea of *A* and *B* in the revival set of the abstract idea of the causal relation. This will, of course, be a revival set that has resulted from the effect on the mind of the specific kind of resemblance that holds among the various object-pairs that, following experience of constant conjunction of like pairs, sustain association and causal inference with its characteristic impression of determination or necessary connection.

Hume aims to capture this revival set with the two 'definitions of "cause"' that follow his account of causal inference:

[1] We may define a cause to be, 'An object precedent and contiguous to another, and where all the objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter'.

[2] 'A cause is an object precedent and contiguous to another, and so united with it that the idea of the one determines the mind to form the idea of the other, and the

> impression of the one to form a more lively idea of the other'. (THN 1. 3. 14. 30–1; SBN 169–70)

These two definitions express what we might call 'external' and 'internal' conditions—constant conjunction and association-plus-inference, respectively—that lead to the impression of necessary connection and the inclusion of an object pair in the revival set of the abstract idea of cause and effect. They thus provide alternative specifications of the content of the revival set—which is why Hume calls them 'definitions', since a definition is, for him, as it was for Locke, simply a specification of the idea for which a term stands. (The interpretation of these two definitions, and of the relation between them, has generated a vast literature, partly because they do not appear to be extensionally equivalent: the first but not the second seems satisfiable in the absence of actual observers, while the second but not the first appears satisfiable by unrepresentative samples. However, I have argued (Garrett 1997: ch. 5) that each definition can be understood in both a primary absolute sense and a secondary subjective sense, and that the two definitions are necessarily coextensive in their absolute senses and again in their subjective senses. The absolute sense of the second definition invokes a hypothetical idealized observer, while the subjective sense of the first definition invokes constant conjunction within the experience of a particular observer. Henceforth, I shall be considering only the primary, absolute sense of each definition.)

Because they all pick out the same revival set, there is a very rough-grained sense in which the two definitions and the term 'cause' itself might all be said to specify 'the same concept'. But in a finer-grained and more natural sense, they clearly do not, for each picks out that common revival set in a different way, so that the act of judging that some pair of objects is constantly conjoined is not the same act as judging that it gives rise to association and inference, and neither of these is the same cognitive act as judging that the objects are cause and effect. Hume's first definition specifies (somewhat vaguely) a concept, CONSTANT CONJUNCTION, that one deploys by means of operations on the revival sets of such other concepts as PRECEDENCY, CONTIGUITY, and RESEMBLANCE, while the second specifies a concept, ASSOCIATION-PLUS-INFERENCE, that one deploys by means of operations on the revival sets of s u c h other concepts as PRECEDENCY, CONTIGUITY, IDEA, IMPRESSION —and (notoriously) DETERMINATION. In fact, one could hardly deploy *either* definition—resolving the first definition's vagueness about what kinds of resemblance are significant or using the second definition's causal concept of DETERMINATION—without already using a separate capacity to detect causal relations independent of the use of these two definitions. In order to deploy the concept of CAUSATION itself, in contrast, it seems that one need deploy *no* other explicit concepts.[3] How is this possible?

In Book 3 of the *Treatise*, Hume discusses distinctive sentiments of 'moral approbation and disapprobation'. These sentiments are—like the impression of necessary connection—internal 'impressions of reflection', and he characterizes our ability to feel them as a 'moral sense' that enables us to discriminate vice and virtue (THN 3. 1. 2). He goes on to offer (in the *Treatise* and even more clearly in *An Enquiry concerning the Principles of Morals*[4]) two definitions of 'virtue', one of which appeals to features external to the moral spectator ('the possession of mental qualities, useful or agreeable to the person himself or to others' (EPM 9.

1; SBN 268)) and one of which appeals to features internal to the moral spectator ('whatever mental action or quality gives to a spectator the pleasing sentiment of approbation' (EPM Appendix 1. 9; SBN 289)). This analogy suggests that the mental operations by which constant conjunction leads to association, inference, and the impression of necessary connection may be viewed as a kind of 'causal sense', one that allows the mind to detect those pairs of objects that are causally related and to distinguish them from those that are not.

Of course, senses need not be, and generally are not, infallible; not all things that appear to resemble each other in a given respect do resemble each other in that respect. Hence, not every object that we might initially include in the revival set of an abstract idea is properly so classified. The sizes, shapes, and relative positions of bodies as they are initially sensed must often be corrected, Hume allows, by consideration of the position of the observer (THN 3. 3. 3. 2; SBN 602–3); and the apparent colours, sounds, tastes, and smells of objects must also sometimes be corrected, he notes, for features of the circumstances of observation, including the health of the sense organs. He emphasizes how the immediate deliverances of the moral sense must likewise be 'corrected' by taking into account differences of perspective on the individuals assessed so as to reduce the 'contradictions' in felt response that the same character would otherwise produce among different observers and even within the same observer at different times (THN 3. 3. 3. 2; SBN 602–3). The initial deliverances of the sense of beauty must often be corrected as well, both by a consideration of the physical circumstances of observation (3. 3. 1. 15; SBN 582) and by reflectively developed rules of criticism (2. 2. 8. 18; SBN 379).

In parallel fashion, Hume offers, in *Treatise* 1. 3. 15, a set of eight 'rules by which to judge of causes and effects', rules that serve to guide the refinement of the revival set of one's abstract idea of the relation of cause and effect. These rules, he reports, 'are formed on the nature of our understanding, and on our experience of its operations in the judgments we form concerning objects'—that is, by reflection on the mechanism of causal inference in the light of the past successes and failures of causal reasonings of various kinds—and by means of them 'we learn to distinguish the accidental circumstances from the efficacious causes' (1. 3. 13. 11; SBN 149). Of course, the greatest problem of perspective or situation in discerning causal relations lies in the limitation of our experience to only a small part of what actually occurs in the world, with the resulting danger of insufficient or unrepresentative samples. Inquiry into causes involves the development and use of both experiments and rules for judging that mitigate this insufficiency as much as possible. Just as the correct or true revival set for the abstract idea of a sensible, moral, or aesthetic quality is that which would arise in an ideal human observer judging in accordance with proper rules of correction, so the correct or true revival set for the abstract idea of cause and effect is the set that would arise in an ideal observer having the human causal sense, possessed of enough observations to constitute a sufficient and representative sample for any causal judgement, and employing the proper rules for judging of causes and effects.

Two additional points are worth noting. First, two different concepts may in fact pick out the same quality (for example, the concepts of RED and SUCH-AND-SUCH SURFACE REFLECTANCE PROPERTY) or relation. If they do pick out the same quality or relation, of course, this will typically be discoverable only a posteriori, even though the identity itself will—like all identities—be necessary. Second, Hume recognizes that, once we have explicit concepts of

relations, we may use them to form what he (like Locke and Berkeley) calls 'relative ideas' of things; for example, an idea of 'the companion of Rousseau' or 'the cause of malaria'. To form such an idea, one need not know any of the specific qualities of the thing itself, so long as one has an idea of the relation and of the other relatum or relata.

4. CAUSAL PROJECTIVISM

We are now in a position to see what Hume would grant to causal projectivists and what he would deny. First, of course, he would readily grant the psychological claim that human causal inference typically involves a projective mislocation of an internal feeling. As a result of this projective mislocation, human beings will typically include ideas of necessary connection as a part of the ideas of cause-and-effect pairs that make up the revival set of the abstract idea of CAUSATION. In this respect, the concept of CAUSATION itself in the minds of those who make the projection will differ from any concept in the minds of individuals who do not. Second, he would also grant, as Simon Blackburn (1990; 2000) maintains, that causal claims *express* a readiness to make inferences with a certain necessary modal character—an attitude that Blackburn calls 'causalizing'. This is one respect in which those who make causal judgements have a psychological state different from the states of individuals who merely register constant conjunctions. On the other hand, Hume would deny at least three claims that are sometimes made by projectivists as a basis for concluding that real causal relations find no place in the metaphysical structure of the universe.

First, he rejects both *non-cognitivism* and *error theory* about ascriptions of causal relations themselves; that is, he holds that causal attributions are strictly apt for evaluation as true or false and that many of them are in fact true. For example, he declares himself in the Introduction to the *Treatise* to be pursuing nothing less than the 'truth' about the 'principles of human nature', a project that requires 'forming a notion' of the 'powers and qualities' of the mind and 'explaining all effects from the simplest and fewest causes'.[5] He clearly aims at least to contribute to the development of a true account of the causal relations exemplified in the human mind.

It may seem that his concessions to projectivism undermine this commitment. For he defines 'truth or falshood' as 'an agreement or disagreement either to the *real* relations of ideas, or to *real* existence and matter of fact' (THN 3. 1. 1. 9; SBN 458; see also THN 2. 3. 3. 5; SBN 415); and has he not admitted that our concept of CAUSATION contains, as an essential part, an element that does not 'agree with' any intrinsic feature of cause-and-effect pairs? In response to this objection, we must clarify what it might mean, for Hume, to say that an idea is 'an essential part' of a concept.

In Locke's theory of ideas (Locke 1979/1689), it is not always clear which ideas are concepts, but it is always a straightforward question whether one concept is a part of another. The concept of *SOLIDITY*, for example, is literally part of the complex concept of *BODY* for Locke, while the idea of *RED* is not. In Hume's theory, by contrast, it is always clear which ideas are concepts—as noted, they are the ones that he calls 'abstract ideas'—but questions about the parts of concepts are not so straightforward. In both these respects, Hume's theory marks an advance over Locke's. In a Humean theory, saying that an idea is 'part of' a concept

might mean any of these:

(1) The idea is an element in the revival set of the concept. For example, the idea of Lassie is part of the revival set of the concept DOG.

(2) The idea is itself an abstract idea that must be used in arriving at the revival set of the concept. For example, the revival set of the concept FEMALE DOG is obtained by taking the intersection of the revival sets of FEMALE and DOG.

(3) The idea is part of every idea in the revival set. For example, an idea of a right angle is a proper part of every idea in the revival set of the concept of RIGHT TRIANGLE.

It is not plausible to suppose that the idea of necessary connection is part of the concept of CAUSATION in the first way for Hume. No doubt some of his readers have supposed that he regards the concept of NECESSARY CONNECTION, along with the concepts of PRIORITY and CONTIGUITY, as a part of the concept of CAUSATION in the second way, but this is a mistake. The standard concept of CAUSATION is acquired through the use and refinement of a causal sense, not through performing combinatory operations on other concepts; it is quite possible that a mind could acquire the explicit concept of CAUSATION before acquiring the explicit concepts required to deploy either of the two definitions. Indeed, it is likely that relatively few concepts signified by single words (as opposed to phrases) have 'parts' in this sense on Hume's view.

It is true, in contrast, that the idea of necessary connection is part of at least many individuals' concept of CAUSATION in the third way. The inclusion of the idea of necessary connection in ideas of members of the revival set presumably facilitates the development of the concept itself, as it gives additional salience to the resemblance among ideas of cause-and-effect pairs beyond their exemplification of constant conjunction and their role in association and inference. However, it is not *essential* to the concept (that is, 'abstract idea') of CAUSATION itself. For to overcome the projective illusion—as Hume implies can be done with care and attention (THN 1. 3. 14. 31, 1. 4. 3. 9–10, and 1. 4. 7. 5–6; SBN 170, 222–4, and 266–7)—is not to change the pairs of objects whose ideas constitute the revival set of the idea of cause and effect, but only to correct the way in which those pairs are represented. In order to represent Boston and New York as being 190 miles apart, it is not necessary to use two ideas that exactly resemble those cities and are themselves 190 miles apart; and the idea of blue, Hume allows, can be used to represent blue objects even if what he calls 'the modern philosophy' is correct in thinking that those objects have no quality literally resembling that idea (Hume 1978*b*/1777, 'Of the Standard of Taste', 233). To include a lively idea of a pair of objects in the revival set of CAUSATION when the pair does indeed 'agree' with the abstract idea by resembling its other members in the manner picked out by the causal sense is, prima facie, to believe truly that the objects are related by cause and effect. The relation *is* in fact a causally necessary one, as he understands causal necessity, regardless of whether the impression of necessary connection is 'spread' onto the 'inseparable' objects or not.

A second projectivist claim that Hume denies is that causation itself is mind-dependent. Hume considers explicitly the objection that his theory of causation would 'reverse the order of nature' by making causes depend on thought, rather than vice versa. In doing so, he

imagines an opponent who exclaims,

What! The efficacy of causes lie in the determination of the mind! As if causes did not operate entirely independent of the mind, and would not continue their operation, even tho' there was no mind existent to contemplate them, or reason concerning them. (THN 1. 3. 14. 26; SBN 167)

Hume's response is as follows:

As to what may be said, that the *operations of nature* are independent of our thought and reasoning, I allow it; and accordingly have observ'd, that objects bear to each other the relations of contiguity and succession; that like objects may be observ'd in several instances to have like relations; and that all this is independent of, and antecedent to the operations of the understanding. (THN 1. 3. 14. 28; SBN 168; emphasis added)

Note that Hume allows that the 'operations of nature' that he grants to be mind-independent are the operations of 'causes'. This is consistent with his other frequent uses of the term 'operation' (ninety occurrences in the *Treatise* alone), which he consistently treats as a causal one. Such an admission makes perfect sense on his view. For the kinds of resemblance among pairs of objects that the refined and idealized causal sense of actual human beings picks out would continue to hold whether there were human beings or not, and events in nature would continue to exemplify the true generalizations of a unified theory of nature whether that theory were formulated or not. It might be argued that cause-and-effect pairs have no resemblance or common feature *except* their ability to stimulate the causal sense; but Hume's fifth 'rule by which to judge of causes and effects' requires that similar effects are always produced by some shared feature of the causes: 'Where several different objects produce the same effect, it must be by means of some quality which we discover to be common amongst them. For as like effects imply like causes, we must always ascribe the causation to the circumstance wherein we discover the resemblance' (THN 1. 3. 15. 7; SBN 174).

Finally, Hume would deny that causal relations are explanatorily dispensable. Simon Blackburn both holds and attributes to Hume a sophisticated 'quasi-realist' form of causal projectivism that readily allows the truth of many causal ascriptions. He nevertheless rejects what he calls 'theoretical or ontological ... [or] upper-case Realism' about causation by employing the following test: to be an 'upper-case Realist' about a kind of entity is to allow that 'we cannot understand what is going on [in discourse about the entity] except in terms of our responding to a world whose entities and properties and relations are the ones ostensibly referred to' in the discourse (Blackburn 2000: 110). Applying that test to the present case, however, it seems evident that, for Hume, no explanation of causal discourse could avoid making reference to causal relations as part of the explanation—by invoking, for example, our *responding to* observed constant conjunctions and the role of such reactions in *producing*

inference and discourse. Without such causal references, it seems, one would have at most a very long narrative about, rather than a genuine explanation of, causal discourse.

5. CAUSAL REDUCTIONISM

What can Hume grant to causal reductionists? First, of course, he can agree that many causal judgements are true. More specifically, he can grant that constant conjunction—or rather, the very specific *kind* of constant conjunction that is in fact detected by the corrected causal sense, with its specific standards of relevant resemblance—is both necessary and sufficient for the existence of a causal relation. Indeed, this constant conjunction is—like the satisfaction of Hume's coextensive second definition of 'cause'—necessary and sufficient for the existence of a causal relation that is itself an example of causal *necessity*. Were this not the case, he could not conclude, as he does in 'Of liberty and necessity' (THN 2. 3. 1 and EHU 8), that the uniform conjunction of particular kinds of human actions with particular kinds of circumstances, characters, and motives, is (like the satisfaction by these things of the coextensive second definition of 'cause') sufficient for the causal necessity of those actions—regardless of other respects in which the activity of minds might differ from that of bodies. Furthermore, he can also grant that the relation of causation might well *be* the relation of this particular kind of constant conjunction, just as the quality of redness may prove to be the quality of having a certain surface reflectance property, and the quality of virtue to be the possession of features of character useful or agreeable to their possessor or others.

But while Hume is friendly to these metaphysical claims of the reductionist, he would deny nearly all of the semantic reductionism that has often supported them. He would deny that either of his two definitions of 'cause' is synonymous with the term 'cause' itself; for under the most natural standard for individuating concepts, they specify three different concepts. (It is just as well, of course, that neither definition is synonymous with 'cause'; Hume treats the definitions as parallel, and they are obviously not synonymous with each other.) In fact, the Humean concept of CAUSATION has no semantic analysis into *conceptual* parts at all. (Similarly, more specific causal concepts such as those expressed by the transitive verbs 'move', 'drop', and 'convince' can be semantically simple and need not have the more general concept CAUSATION as a part.) He would also deny the related epistemological claim that knowledge of constant conjunction is alone sufficient for knowledge of causal relations—for in addition, one must know, a posteriori, that constant conjunction is what the causal sense detects.

Finally, Hume has reason to be sceptical of a broad and rather vague metaphysical claim often made by (and even more often ascribed to) reductionists: namely, that there is definitely 'nothing more to causation than constant conjunction'. Of course, if this means simply that causation is the same relation as a particular kind of constant conjunction, then he can grant that it *may* well be true. But even if that is true, there may be *additional* features of cause-and-effect pairs that are highly relevant to their being so related—just as there might be underlying subatomic features of bodies in virtue of which they have given surface reflectance properties, or features of human physiology intimately tied up with having features of character that are useful or agreeable to their possessor or others. Indeed, being so related as to produce

association and inference in a corrected human causal sense is already one additional important feature of cause-and-effect pairs, and there may well be others of various kinds—including some of these postulated by contemporary theories of causation.

6. CAUSAL REALISM

Like causal realists, Hume can accept causal cognitivism while rejecting the semantic reductionism that treats causal claims as synonymous with claims about constant conjunction. Moreover, like causal realists, he is not wedded to the deflationary metaphysical claim that there is 'nothing more' to causation than constant conjunction. Furthermore, he can grant our capacity, via the use of relative ideas, to *entertain the hypothesis* that there is some kind of relation that is neither the intuitive-or-demonstrative necessity of relations of ideas nor the constant-conjunction-based necessity of causes but yet pertains to all cause-and-effect pairs and would, if grasped, be judged to be a *third* kind of necessity. He can also grant—what is something yet different—the capacity to wonder whether our present cognitive faculties are so deficient that more adequate ideas of things would reveal the truth of the Spinozistic hypothesis that all causes and effects are actually related by the intuitive or demonstrative necessity of relations of ideas. Indeed, the character of Cleanthes in Hume's *Dialogues Concerning Natural Religion* appears to consider a parallel hypothesis concerning the alleged 'necessary existence' of God:

It is pretended that the Deity is a necessarily existent being; and this necessity of his existence is attempted to be explained by asserting, that if we knew his whole essence or nature, we should perceive it to be as impossible for him not to exist, as for twice two not to be four. But it is evident that this can never happen, while our faculties remain the same as at present. It will still be possible for us, at any time, to conceive the non-existence of what we formerly conceived to exist; nor can the mind ever lie under a necessity of supposing any object to remain always in being; in the same manner as we lie under a necessity of always conceiving twice two to be four. The words, therefore, necessary existence, have no meaning; or, which is the same thing, none that is consistent.

But further, why may not the material universe be the necessarily existent being, according to this pretended explication of necessity? We dare not affirm that we know all the qualities of matter; and for aught we can determine, it may contain some qualities, which, were they known, would make its non-existence appear as great a contradiction as that twice two is five. . . . It must be some unknown, inconceivable qualities, which can make his non-existence appear impossible, or his attributes unalterable: and no reason can be assigned, why these qualities may not belong to matter. As they are altogether unknown and inconceivable, they can never be proved incompatible with it. (Hume 1935/1779: 9. 6–7)

At the same time, however, Hume would reject the suggestion that two objects exhibiting the kind of constant conjunction (and hence causal necessity) that is detected by a corrected causal sense could nevertheless *fail* to exhibit a genuine causal relation through failure to possess some *other* kind of necessity—just as he insists elsewhere that the corrected moral

sense is the ultimate authority concerning vice and virtue (THN 3. 2. 9. 4; SBN 522; and EHU 8. 35; SBN 102–3). Moreover, he would reject the proposal of Galen Strawson (1989) that we *need* to infer the existence of such further necessity in order to explain the uniformity of nature. For each individual event is sufficiently explained by the laws of nature that it exemplifies; there is, for Hume, no more need for a further 'necessary cause of uniformity' than there is a need for a necessary first cause of the universe as a whole when every event within it already has its own cause. As Cleanthes continues in the passage just cited:

In such a chain, too, or succession of objects, each part is caused by that which preceded it, and causes that which succeeds it. Where then is the difficulty? But the WHOLE, you say, wants a cause. I answer, that the uniting of these parts into a whole, like the uniting of several distinct countries into one kingdom, or several distinct members into one body, is performed merely by an arbitrary act of the mind, and has no influence on the nature of things. Did I shew you the particular causes of each individual in a collection of twenty particles of matter, I should think it very unreasonable, should you afterwards ask me, what was the cause of the whole twenty. This is sufficiently explained in explaining the cause of the parts. (1935/1779: 9. 11)

Furthermore, Hume would deny that we generally *do* use relative ideas to postulate additional unknowable necessitating powers in causes or necessitating principles governing them when we think about causation; instead, we typically engage in a correctable projective error that results in conflating two *known* and distinct kinds of necessity in a way that results in pronouncements about power or causal necessity (that is, necessity of the second kind) that are, as he says, impossible or meaningless. This is because we treat a *single* necessity as both causal in the ordinary known way (and so in fact consisting in a psychological inseparability derived from experienced constant conjunction) and yet intrinsic to the causes and effects themselves. While he sometimes characterizes power or necessity as described by such confused pronouncements as 'unknowable', this does not show that he does not *also* regard it as contradictory, impossible, or (informally) 'meaningless'. Indeed, the passage quoted above from *Dialogues* 9.6–7 shows Cleanthes characterizing God's necessary existence (as proposed apart from an unknown necessity conceived relationally through a speculative change in our faculties) first as meaningless or contradictory and then also as unknowable. Like Berkeley, he appears to regard the unknowable as a species of the meaningless or contradictory.

The postulation of intrinsic powers exhibiting a third and unknown kind of necessity, on the other hand, requires a *doubly* relational idea: for the intrinsic power is conceived as an otherwise unknown *something* in causes that grounds a necessitation relation that is equally unknown except through its postulated relation of resemblance to known kinds of necessity. While the use of such a doubly relational idea is certainly possible, Hume gives little genuine indication that he himself uses such an idea, nor that he believes that something answers to it, nor that he sees epistemic merit in the hypothesis that there is something answering to it. However, he could certainly explain the psychology of someone who did do these things—and they could still be epistemic allies on a wide range of fronts.

7. CONCLUSION

What then, according to Hume, is the causal relation? It is, as he promises at the outset of his exploration of it, the relation that is 'discovered' by the process of causal inference—a process that we may therefore regard as the basis of a 'causal sense', whereby objects of types that are constantly conjoined produce association, inference, and an impression of determination or necessary connection. It is by means of this sense and the judgements based upon it that we discriminate cause-and-effect pairs from other pairs, in something like the way that Hume's visual sense discriminates colours, his moral sense discriminates vice and virtue, and his aesthetic sense discriminates beauty and deformity.

In the course of explaining this causal sense and the causal judgements to which it gives rise, Hume asserts that it involves the projection onto cause-and-effect pairs of an element of felt necessity that does not resemble anything in those pairs, deriving its origin instead entirely from the experienced constant conjunction of types of objects. Yet at the same time, he regards many causal judgements as true, mind-independent, and even explanatorily basic. He grants that constant conjunction of the kind ideally detected by the causal sense is sufficient for causation. Yet he does not suppose that ascriptions of causation are synonymous with ascriptions of constant conjunction or known solely on their basis, and he does not insist that there is nothing more to causation than constant conjunction. He rejects the possibility of straightforwardly conceiving—and hence, too, of straightforwardly believing—that there are *inherently* necessitating causal powers or ultimate necessitating causal principles in causes themselves. Yet he does not positively rule out the prospect that some by-us-inconceivable analogues (analogues through a by-us-inconceivable resemblance) of such powers or principles may characterize causes and effects. His causal sense theory allows him to do all these things with complete consistency.

Should we classify Hume's resulting theory of causation as projectivism, reductionism, or realism? If the interpretation offered here is right, then he concedes something to the motivations of each of these packages, and he could with some justice be classified as subscribing to any of them, or all of them, or none or them—depending on the details of the more specific definitions that might be proposed for them.

FURTHER READING

Projectivist interpretations include Stroud (1977; 1993), and Pears (1990); Blackburn (1990; 2000) and Conventry (2006) offer sophisticated 'quasi-realist' projectivist interpretations. Beebee (2006) is also sympathetic to projectivist interpretations.

Examples of reductionist interpretations—setting out what is often called simply the 'Humean view' of causation—include Robinson (1962), Ayer (1973), Woolhouse (1988), and Clatterbaugh (1999).

Recent realist interpretations include Broughton (1987), Wright (1983; 2000), Craig (1987) (but see Craig 2000), and Strawson (1989; 2000). Winkler (1991) provides an influential and textually detailed response. See also Costa (1989); Buckle (2001); and Kail (2001; 2003; 2007; 2008). Kail, following a suggestion of Craig (2000), treats causal projectivism as a purely psychological doctrine and causal realism as a semantic doctrine in interpreting Hume.

For an overall view of the debate, see Read and Richman's *The New Hume Debate* (2000, or

the revised 2008 edn.).

REFERENCES

AYER, A. J.(1973). *The Central Problems of Philosophy*. Harmondsworth: Penguin.

BEEBEE, H. (2006). *Hume on Causation*. London: Routledge.

BLACKBURN, S. (1990). 'Hume and Thick Connexions'. *Philosophy and Phenomenological Research* 50 (supplement): 237–50. (Reprinted in Read and Richman 2000.)

—— (2000). Postscript to 'Hume and Thick Connexions', in Read and Richman 2000: 100–12.

BROUGHTON, J. (1987). 'Hume's Ideas about Necessary Connection', *Hume Studies* 13: 217–44.

BUCKLE, S. (2001). *Hume's Enlightenment Tract: The Unity and Purpose of* An Enquiry Concerning Human Understanding. Oxford: Clarendon.

CLATTERBAUGH, K. (1999). *The Causation Debate in Modern Philosophy:* 1637–1739. New York: Routledge.

COSTA, M. (1989). 'Hume and Causal Realism'. *Australasian Journal of Philosophy* 67: 172–90.

COVENTRY, A. (2006). *Hume's Theory of Causation: A Quasi-Realist Interpretation*. London: Continuum.

CRAIG, E. (1987). *The Mind of God and the Works of Man*. Oxford: Clarendon.

—— (2000). 'Hume on Causality: Projectivist *and* Realist?', in Read and Richman 2000: 113–21.

GARRETT, D. (1997). *Cognition and Commitment in Hume's Philosophy*. New York: Oxford University Press.

—— (forthcoming). 'Hume's Theory of Causation: Inference, Judgment, and the Causal Sense', in D. C. Ainslie (ed.), *The Cambridge Companion to Hume's Treatise*. Cambridge: Cambridge University Press.

HUME, D. (1935/1779). *Dialogues concerning Natural Religion*, ed. Norman Kemp Smith. Oxford: Oxford University Press.

—— (1975/1748 and 1751). *Enquiries concerning Human Understanding and the Principles of Morals*, ed. L. A. Selby-Bigge, rev. P. H. Nidditch. Oxford: Clarendon.

—— (1978a/1739–40). *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, rev. P. H. Nidditch. Oxford: Clarendon.

—— (1978b/1777). *Essays Moral, Political, and Literary*, ed. E. F. Miller. Indianapolis: Liberty Fund.

—— (1998/1751). *An Enquiry concerning the Principles of Morals*, ed. T. L. Beauchamp. The Clarendon Edition of the Works of David Hume. Oxford: Oxford University Press.

—— (2000a/1739–40). *A Treatise of Human Nature,* ed. D. F. Norton and M. J. Norton. Oxford Philosophical Texts. Oxford: Oxford University Press.

—— (2000b/1748). *An Enquiry concerning Human Understanding*, ed. T. L. Beauchamp. The Clarendon Edition of the Works of David Hume. Oxford: Oxford University Press.

KAIL, P. J. E.(2001). 'Projection and Necessity in Hume'. *European Journal of*

Philosophy 9: 24–54.

—— (2003). 'Is Hume a Causal Realist?'. *British Journal for the History of Philosophy* 11: 509–20.

—— (2007). *Projection and Realism in Hume's Philosophy* . Oxford: Oxford University Press.

—— (2008). 'Is Hume a Realist or an Anti-Realist?', in E. Radcliffe (ed.), *The Blackwell Companion to Hume*. Oxford: Blackwell.

LOCKE, J. (1979/1689). *An Essay concerning Human Understanding,* ed. P. H. Nidditch. Oxford: Oxford University Press.

PEARS, D. (1990). *Hume's System: An Examination of the First Book of his Treatise* . Oxford: Oxford University Press.

READ, R., and RICHMAN, K. A. (eds.) (2000). *The New Hume Debate*. London: Routledge. Rev. edn. 2008.

ROBINSON, J. A.(1962). 'Hume's Two Definitions of "Cause"'. *The Philosophical Quarterly* 12: 162–71.

STRAWSON, G. (1989). *The Secret Connexion: Causation, Realism, and David Hume*. Oxford: Clarendon.

—— (2000). 'David Hume: Objects and Power', in Read and Richman 2000: 31–51.

STROUD, B. (1977). *Hume*. London: Routledge & Kegan Paul.

—— (1993). '"Gilding" and "Staining" the World with "Sentiments" and "Phantasms"'. *Hume Studies* 19: 253–72. Repr. Read and Richman 2000.

WINKLER, K. (1991). 'The New Hume'. *The Philosophical Review* 100: 541–79. Repr. Read and Richman 2000.

WOOLHOUSE, R. (1988). *The Empiricists*. Oxford: Oxford University Press.

WRIGHT, J. P. (1983). *Hume's Skeptical Realism* . Minneapolis: University of Minnesota Press.

—— (2000). 'Hume's Causal Realism: Recovering a Traditional Interpretation', in Read and Richman 2000: 88–99.

CHAPTER 5
KANT

ERIC WATKINS

1. INTRODUCTION

Kant's views on causality have long been of interest to philosophers as promising an alternative to Hume's empiricist account without thereby falling back into a strictly or straightforwardly rationalist position. Slightly more specifically, Kant holds that a causal principle according to which every event has a cause, or follows according to a causal law, cannot be established through induction as a purely empirical claim, since it would then lack strict universality, or necessity. Nor can it be justified through a mere analysis of what is involved in the principle of sufficient reason, since the assertion of the principle of sufficient reason itself can appear to be dogmatic, or unwarranted. Instead it must be warranted by appealing to the very possibility of experience, that is, through the idea that such a principle must be presupposed if we are to have any experience at all, or at least, experience of a distinctively human kind. This kind of justification is often called 'transcendental' since it is based on the conditions of the possibility of experience of things as contrasted with the strictly 'metaphysical', which would derive solely from the conditions of the possibility of things. It is then traditionally linked, following Kant's own provocative meta-philosophical reflections, to asserting that this causal principle would be neither a synthetic a posteriori proposition (as the empiricists might suggest), nor an analytic a priori proposition (as the rationalists would have it), but rather a synthetic a priori proposition, that is a substantive or informative proposition about the world that can be known to hold independently of any particular experience, or empirical evidence, we might have. In this way, one can see how causality forms the cornerstone for Kant's Critical philosophy as a whole.

In the following I briefly describe the historical context in which Kant develops his account of causality, and then clarify some central features pertaining to the meaning, justification, and presuppositions of the claims that form the heart of this account before concluding with a brief sketch of how his views on causality are incorporated into his views in physics and biology. Accordingly, I discuss Kant's views on causality in (2) his pre-Critical period (i.e. prior to the publication of the *Critique of Pure Reason* in 1781), (3) the *Critique of Pure Reason* (especially (3.2), the Second and (3.3) Third Analogies of Experience), (4.1) the *Metaphysical Foundations of Natural Science*, and (4.2) the *Critique of the Power of Judgment*.

2. THE PRE-CRITICAL PERIOD

Causality was a topic of fundamental interest to Kant from the very beginning of his career. He began to develop his most basic account of causality in the 1750s, most prominently and explicitly in the *Nova Dilucidatio*, by way of contrast with the views of Leibniz, Wolff, and their successors in Germany. However, by the early 1760s he seems to have become aware of Hume's position, and modified his view in the *Negative Magnitudes* essay so as to respond to the challenge he saw it as posing for his own account. By the time of his appointment, in 1770, to a professorship of philosophy at the university in Königsberg, he came to view causal relations as what binds substances together to form a single world so that in his Inaugural Dissertation causality is nothing less than the most fundamental principle of the form of the world.

2.1 The *Nova Dilucidatio:* The Initial Response to Leibniz, Wolff, and Crusius

Kant's primary aim in the *Nova Dilucidatio* (1755) is to argue for two distinctive causal principles, the Principle of Succession and the Principle of Coexistence, which, despite considerable revisions along the way, form a significant component of his account of causality throughout his career. The Principle of Succession states that substances can change their intrinsic states only if (a) they stand in reciprocal causal relations to each other and (b) the relations between these substances are themselves changing, since the grounds that are posited along with any given substance are unchanging and thus unable to account for change. The Principle of Coexistence maintains that only the divine understanding can enable the reciprocal interaction asserted by the Principle of Succession, since substances do not stand in relations to each other (whether causal or otherwise) by means of their 'mere existence'.[1] (For if constrained only by the conditions necessary for their existence, two substances could exist in different spatial relations; the ground of whatever spatial relations they have to each other cannot therefore be given by their mere existence.) The Principle of Succession signals Kant's fundamental departure from Leibniz's and Wolff's Pre-established Harmony, which holds that a finite substance can act causally only on itself, not on others, while the Principle of Coexistence reveals Kant's rejection of one of the most prominent anti-Leibnizian positions in Germany at the time developed by Christian August Crusius, according to which the mere existence of one substance can be causally responsible (as an 'existential ground') for effects in another.

2.2 The Essay on *Negative Magnitudes:* The Initial Response to Hume

Hume's *Enquiry concerning Human Understanding* was translated into German in 1755, and it is clear that Kant had become familiar with at least some aspects of Hume's position by the early 1760s. For in his *Attempt to Introduce the Concept of Negative Magnitudes into Philosophy*, published in 1763, Kant modifies the account of causality he had developed in the *Nova Dilucidatio* in response to the challenge that Hume's position presented for it. Specifically, Hume argues in an especially clear and forceful way that since reason does not

allow an inference from the existence of one thing to the existence of another, we cannot assert necessary connections between such things. However, this is precisely the position Kant had been affirming against Leibniz and Crusius in asserting that substances that undergo changes in their states are necessarily connected to each other by means of their causal relations. He responds to this challenge by drawing a distinction between logical and real grounds, where logical grounds are rational principles based on the 'principle of identity', or the principle of contradiction, and by asserting that causal relations are based on real rather than logical grounds. As a result, he can then agree with the Humean point that reason alone cannot infer the existence of one substance from the existence of another—causal relations between them are not logical relations ascertainable by reason—while still maintaining that causality involves a necessary connection between substances. At this point, however, Kant explicitly recognized that he had no positive account of what principle real grounds are based on (if not the principle of identity), what kind of necessity real grounds support (if not logical necessity), and what faculty might be able to cognize real grounds and the relations they support (if not reason).

2.3 The Inaugural Dissertation

In his Inaugural Dissertation Kant famously distinguishes between the sensible and intelligible worlds, attempts to describe the distinctive principles of each, and so comes to the position that space and time are merely subjective principles that pertain solely to the form of the sensible world, a position he would expand on in the *Critique of Pure Reason*, resulting in the rich and controversial metaphysical and epistemological theses of Transcendental Idealism. However, as a preliminary to his discussions of the concepts of the sensible and intelligible world, he introduces and argues for a generic concept of the world. Every world must have both a matter—a material out of which it is made—and a form—a principle of organization for that matter. While Kant, following tradition, identifies substance as the matter of any world, he asserts that its form consists of mutual interaction, since given a plurality of substances, a principle is needed that can unite them into a *single* world and so that each substance is not (as Leibniz had maintained) a world unto itself, existing apart from all others. This account of the world is significant for Kant's development because it illustrates how he wants to deploy causality to a specific, systematically crucial purpose, namely that of explaining the underlying unity of the world.

3. The Critical Period

Causality continued to be an important topic for Kant in the *Critique of Pure Reason*, as his reflections on the meaning and justification of the concept of causality led him to develop novel components of his epistemology (e.g. the pure concepts of the understanding, or categories), and his 'critical turn', with its rejection of purely metaphysical arguments as dogmatic, motivated him to formulate new transcendental arguments for causality and mutual interaction in the Second and Third Analogies of Experience.

3.1 The Categories of Causality and Mutual Interaction

Kant's idea that causality is to be understood in terms of real grounds led him to provide an account of the meaning of the concept of causality that differs radically from Hume's empiricist account. First, according to Kant, our most primitive notion of causality is not an empirical concept, drawn straight from experience, but it is also not a purely logical concept. Instead, it is derived, he now thinks, by means of a single principle from the unity of the pure understanding (A65/B89), along with other primitive non-empirical concepts such as that of mutual interaction and substance. As a result, it contains both a meaning and a kind of necessity that Hume's concept of causality lacks. Second, Kant stresses similarities between the concept of causality and the hypothetical form of judgement, since both express a kind of conditionality, or dependence, of one item on another, even if hypothetical judgements and causal concepts express dependence between different kinds of things (propositions versus objects) and different kinds of dependencies (logical as opposed to transcendental). Analogous points hold for the category of mutual interaction and the disjunctive form of judgement, in so far as mutual interaction expresses a kind of reciprocal dependence that is similar in certain limited respects to the kind of relationship that holds between those members of a disjunctive judgement that exclude each other in exhausting a certain conceptual space.

Since the categories of causality and mutual interaction are derived from a non-empirical source—the pure understanding—and thus have a non-empirical content, but must still, Kant argues, be applied to empirical objects, he recognizes that they must be given 'schemata', or temporal meanings for their application to be possible. Kant states (without argument) that the schema of causality is the succession of the states of a substance in so far as it is subject to a rule, whereas that of mutual interaction is the coexistence of the states of two substances according to a rule involving their reciprocal grounding. As a result, Kant continues to maintain a link between causality and mutual interaction, on the one hand, and succession and coexistence, on the other, though he now maintains that these causal notions make possible not only a single world, but also a single experience of that world (in the form of the 'unity of experience' or 'one experience').

3.2 The Second Analogy of Experience

3.2.1 *The Claim*

The claim of the Second Analogy of Experience states that causality is a condition of the possibility of our experience of succession. This claim represents a crucial part of Kant's larger project in the *Critique of Pure Reason* in so far as it is an especially important instance of his more general claim that experience, in this case temporal experience, is possible only if the categories, in this case that of causality, are applicable to what is given to us through the senses. Since Hume denied that there are any non-empirical concepts and was, at least on certain interpretations, sceptical about any notion of causality that would contain what Kant

calls 'objective necessity', the Second Analogy has been taken by many to form the crux of Kant's reply to Hume and to empiricism as well. It should be immediately noted, however, that regardless of how one understands Kant's reply to Hume, the Second Analogy does not attempt to find an impression of causality that would either clarify the specific meaning of causality (so as to distinguish it from 'constant conjunction')[2] or serve as a criterion for identifying instances of cause–effect relationships (versus what are merely accidental correlations). Although Kant does mention the irreversibility of the order of our representations when we perceive an instance of objective succession, this need not mean that it functions as a basic premiss in his argument.

Before turning to the argument of the Second Analogy, it is helpful to keep in mind that the very claim of the Second Analogy has been subject to quite different interpretations. The standard view, held by Strawson (1966), Buchdahl (1969), Beck (1978), and Allison (1982), has been to interpret it as asserting that every event has a cause, an assertion whose justification Hume had called into question in the *Treatise*. More recently, however, it has been argued, in very different ways by Guyer (1987), Friedman (1992*a*), and Watkins (2005), that Kant is in fact committed to a different and stronger claim, namely that every event is subject to a causal *law*, a claim that Hume had attacked in both the first *Enquiry* and the *Treatise*. Defenders of the stronger claim have solid textual evidence on their side, since Kant repeatedly refers to rules and on several occasions uses the words 'inevitably' and 'always' to describe how a cause brings about its effect. However, the weaker claim, whose justification is rather controversial in its own right, has seemed more tractable philosophically, since justifying the latter claim is tantamount to the extremely ambitious task of solving the problem of induction.

3.2.2 *The Argument*

While there is widespread agreement that Kant's argument in the Second Analogy takes the form of a transcendental argument—causality in one form or another is supposed to be a condition of the possibility of our experience of succession—there is considerable disagreement about how the argument is supposed to proceed. A first issue concerns the exact meaning of 'experience of succession' and thus the nature of the argument's presupposition and, correspondingly, how broad the scope of its conclusion is supposed to be. One natural temptation is to identify 'experience of succession' with some type of immediate awareness, for example, the succession of subjective representations in the continuous stream of our conscious mental states. Such an interpretation then naturally maintains that the argument is designed to refute the sceptic, since even a radical sceptic must grant the existence of changes in (his or her own) subjective mental states. Both contemporary philosophers and Kant interpreters, however, have hotly contested whether such an argument can succeed. For even if one grants that the concept of succession presupposes the concept of causality, that is, that there is a certain conceptual relation between succession and causality, the radical sceptic can still deny that this conceptual relation entails the existence of anything in the world external to one's thought.[3]

A second line of interpretation, by contrast, interprets experience of succession in the sense of putative objective knowledge, or knowledge of the successive states of an object, and views the argument as showing that such knowledge is possible only by means of the application of the category of causality, since causality is required to determine the states of an object as successive. While this kind of argument poses no threat to the sceptic, given that it presupposes something the sceptic will deny, it is still not a trivial result, since one might doubt, as Hume had, that knowledge of succession requires any causal presuppositions at all. According to Hume, all that is required is that one have first one impression, then another. To show that Hume is wrong about such a claim would thus still be a significant accomplishment.

Regardless of the exact nature of the argument's presupposition and conclusion, one might wonder what drives the argument. What exactly is it, one might reasonably ask, that requires the connection between succession and causality? One prominent strategy, articulated by Bird, has been to investigate the concept of an event in the hope that the concept of causality is somehow implicitly embedded within it. The general idea here is that if an event is defined as a change in an object from one state to another, this notion must contain a rule that determines the order of these states, since the event would not be the very event that it is if these states were reversed. The argument then asserts that whatever it is that determines the order of these states is what we mean by a cause. Thus, the concept of causality is required by the very notion of an event, and therefore for our experience of an event. However, many have been sceptical that the mere concept of an event really analytically includes the notion of a *causal* rule.

A second strategy, articulated especially well by Guyer, asserts that causality is required to solve the problem of time-determination. The problem of time-determination arises because, as Kant repeatedly remarks, we cannot perceive 'time itself', that is, the objective temporal indices at which objects exemplify their different properties. In particular, one must distinguish between 'subjective' and 'objective' time, that is, between the times at which I apprehend a given object and the times at which the object has certain properties. For it would be a mistake, as Kant's example of the ship and the house illustrates, to infer immediately from a succession in our representations to a succession of states in the object itself. (Although I perceive different parts of a house in succession, it would be wrong to infer that the different parts of the house existed only in that order, since they clearly exist simultaneously.) On this line of interpretation, the Second Analogy proceeds by arguing that since the subjective order of our representations by itself cannot account for our knowledge of succession in the object, one must appeal to our knowledge of causality to explain why we think that the states of an object occur in the order in which they do. However, even if knowledge of causality is *sufficient* for knowledge of succession, what is required is that causality (or knowledge thereof, on Guyer's interpretation) be *necessary* for such knowledge —otherwise Hume's position can remain untouched—and it is not clear that this point has been established, since one could, at least in principle, derive knowledge of succession from either *non-causal* principles or from something less than *knowledge* of causal principles.

One can avoid these difficulties by linking succession and causality in the following formal reconstruction of the argument of the Second Analogy:

(P1) Apprehension of objects (the subjective order of perceptions) is always successive.

(P2) There is a distinction between the subjective order of perceptions and the successive states of an object such that no immediate inference from the former to the latter is possible.

(C1) One cannot immediately infer objective succession from the successive order of perceptions. (from P1 and P2)

(P3) To have knowledge of objective succession, the object's states must be subject to a rule that determines them as successive.

(P4) Any rule that determines objective succession must include a relation of condition to conditioned, i.e. that of the causal dependence of successive states upon a cause.

(C2) To have knowledge of the successive states of an object, the object's successive states must be dependent upon a cause, i.e. must stand under a causal rule (from P3, P4, and C1).

The first step of the argument, embodied in P1, P2, and C1, is based on the problem of time-determination and is, I take it, relatively uncontroversial (though Hume objects to a principle similar to P2 in the *Treatise*). It differs from Bird's interpretation in so far as its focus is not on a mere analysis of the concept of an event, but rather on what kind of immediate knowledge we do and do not have. The second step of the argument, constituted by P3, P4, and C2, forges the link between knowledge of succession and causality (rather than knowledge of causality, as Guyer's interpretation holds) and is the most controversial step of the argument—even if one brackets, as this reconstruction implicitly has, considerations about whether the rule referred to in P3 is general, and thus a law, or not. For one might either reject the idea that knowledge of objective succession requires any rule, or admit that it does require a rule, but deny that the rule must be causal.

However, in defence of this step one can note that it is quite natural to ask why a certain state of an object obtains at a certain moment in time (especially when it did not obtain previously), and it is certainly plausible, prima facia, to respond that something caused it to be in that state at that time. In fact, in the contemporary debate concerning free will and determinism, citing anything short of a cause in response to such a question is often thought to constitute an inadequate answer. As a result, it is intuitively plausible, at least at first glance, to assert that only reference to a cause can explain why we should grant that an object has changed its state.

At the same time, intuitions can differ on such a point. The fundamental issue here would seem to be the following. Empiricists, such as Hume, are willing to accept that an object exists in one state at one moment in time and in a different state at a later moment in time, as a thoroughly contingent fact that is revealed to us through our sensory impressions, but not explained with any sort of necessity by rational means. In short, the senses can reveal only *that* something is the case and not *why* it is so. Kant, by contrast, holds that there must be a reason for any change of state we could properly be said to know and attributions of causality

are necessary responses to this situation. For if there were no reason for the change, one would not have an adequate justification for claiming that it occurred and there would also be no reason to view the second state as temporally related to the first state, that is, as its successor state. That is, if there are to be temporal relations that unify the succession states of any object we could experience, then there must be an adequate explanation of them and causal relations are, Kant argues, what is required for this purpose.

3.3 The Third Analogy of Experience

3.3.1 *The Claim and the Argument*

The claim of the Third Analogy runs, at least at a certain level of generality, parallel to that of the Second Analogy, for just as the Second Analogy asserts that causality is necessary for experience of succession, the Third Analogy states that mutual interaction is necessary for experience of coexistence. That is, both Analogies assert that a category is necessary for experience of a certain 'mode of time' or kind of temporal relation. As a result, it is plausible to assume that the assumptions and goals of the Third Analogy—whatever they are—are parallel to those of the Second Analogy. Moreover, the argument of the Third Analogy can be formally reconstructed parallel to that of the Second Analogy:

(P1) Apprehension of substances (the subjective order of perceptions) is always successive.

(P2) There is a distinction between the subjective order of perceptions and the temporal relations (of the states) of substances.

(C1) One cannot immediately infer objective coexistence from the successive order of perceptions (from P1 and P2).

(P3) To have knowledge of objective coexistence, the substances' states must be subject to a rule that determines their states as coexistent.

(P4) Any rule that characterizes objective coexistence must include reciprocally conditioned conditions, that is, a relation of mutual interaction.

(C2) To have knowledge of objective coexistence, substances must stand in mutual interaction (from C1, P3, and P4).

Given these parallels, what is surprising is that the Third Analogy has found far fewer defenders than has the Second Analogy.

At least part of the reason for this disanalogy is that scholars have typically not considered in detail precisely what is involved in the notion of mutual interaction and how it extends beyond the more familiar notion of causality. As a result, the most common reaction to the Third Analogy has been not to object to the fundamental structure of the argument—which would seem to call into question the validity of the argument of the Second Analogy along

with it—but rather to suggest that nothing as robust as mutual interaction is in fact required by knowledge of coexistence. If causality can account for knowledge of succession, then it ought to be able to do so for knowledge of coexistence as well, given the parallels between these two modes of time.

However, one might attempt to defend Kant from this objection by drawing attention to the differences between succession and coexistence, since these differences could make clear why a more robust notion of causality is required to account for coexistence. In this context, two features are noteworthy. First, coexistence is a symmetrical temporal relation, whereas succession is not. Second, as Kant employs these notions, coexistence involves two substances, whereas succession need not (that is, succession is typically the change of state of a single substance). In the light of these differences, it can seem more plausible to assert that coexistence requires mutual interaction, since mutual interaction, as a reciprocal causal relationship between two substances, involves a symmetrical relation between two distinct substances while causality does not. Thus only mutual interaction has the requisite structural features to account for coexistence.

A further cause for discomfort with the Third Analogy arises from a difficulty one might have with the very idea of mutual interaction, at least on what might seem to be the most natural interpretation of it in terms of events. How can two events stand in mutual interaction if mutual interaction is to be understood as Kant does, namely as reciprocal causal dependence? For if a first event causes a second event, then the second event cannot, it would seem, also be the cause of the first event, not if causality entails (as it does for Kant) the claim that the existence of the effect depends on the existence of the cause. That is, mutual interaction would seem to require that each event be both the cause and the effect of the other, which is contradictory (or at least circular).

This difficulty can be resolved by enriching one's ontology so as to include substances and their activities and states and by drawing a distinction between the causal activity of a substance and the (change of) state that is brought about by that activity. For with these resources one can say that in mutual interaction both substances are active causally, without maintaining that they cause each other's existence, since what they cause are changes in each other's state. While this resolution is not available to Hume and to those empiricists who follow him in accepting an ontology consisting exclusively of events, it is clear from Kant's First Analogy of Experience that he views substances and their activities as perfectly acceptable ontological entities, and it is appropriate and even quite natural for the Third Analogy to draw on them.[4]

3.3.2 *The Metaphysics of Causality*

By reflecting on what is required for mutual interaction to represent a coherent possibility, one can better appreciate how it is open to Kant to go beyond what empiricists will acknowledge when he asserts that causality involves irreducible and necessary relations. For if one accepts a Humean (or Lewisian) ontology of events, then it can seem as if (a) causal relations must be reducible to events, since events are the basic building blocks out of which everything else is constructed, and (b) there can be no necessary connections between events, given that events are defined as ontologically distinct states of affairs. However, if one accepts

a more robust ontology, as Kant does, then the accounts of causal relations that are available can be correspondingly richer. As a result, one need not be committed to a reductive analysis of causality (and thus need not be troubled should attempts at reduction encounter difficulties), since causal relations can be a primitive kind of ontological bond, in line with the fact that the category of causality is a primitive concept of the pure understanding, and one can also assert necessary connections in nature on similar grounds.

However, the model of causality that Kant in fact adopts illustrates these abstract assertions in such a way that one can understand the origin of the kind of necessary connections that he claims are involved with causality. For not only can a substance act so as to determine a change of state in another substance, it can do so in accordance with its nature—on the basis of what is essential to it being what it is—a position that Kant had in mind when he introduced 'real grounds' in his pre-Critical period. For example, when one billiard ball acts on another in a collision, its mass, or the amount of substance that it is—an essential feature of matter—is relevant to what effects it can bring about (as are, of course, other circumstances as well). But if a substance acts in accordance with its nature, one can understand how the necessity of causal relations is not a free-floating modality with dubious origins, but rather is grounded in the very essence of the thing that is responsible for such causal relations, which should remove at least some of the mystery that can seem to attach to Kant's affirmations of necessary connections.[5] Not only do substances bring about successive states (such that the activities of substances provide the reason for the succession of the states, which need not therefore be accepted as brute facts), but if they act in accordance with their nature (e.g. are constrained by their nature), what they accomplish can have an element of necessity as well (without thereby denying aspects of contingency).

Moreover, the richer metaphysical resources involved in Kant's model of causality can also support a more robust account of causal laws. It should come as no surprise that Kant would not be especially tempted by a thoroughly empiricist account of causal laws in terms of regularities, but that leaves unspecified what his positive account might be. Because his model of causality invokes substances acting in according with their natures, he can maintain that the laws of nature are based on the natures of things. While such a position is similar in some respects to necessitarian views developed recently by Armstrong (1983) and Tooley (1977), the fact that it is tied so closely to his model of causality and is embedded within a transcendental context can help him to avoid some of the objections that have been raised against aspects of certain necessitarian positions.[6]

4. CAUSALITY AND THE SCIENCES

While it is clear that Kant devoted considerable effort to establishing that causality and mutual interaction are required for the experience of succession and coexistence and thus for the unity of our experience of the world, it is equally evident that his interests in causality extend beyond this central case to how it is employed in particular sciences, such as physics and biology.

4.1 The *Metaphysical Foundations of Natural Science:* Physics

In the *Metaphysical Foundations of Natural Science* (1786), Kant attempted to apply the transcendental principles articulated in the *Critique of Pure Reason* to the empirical concept of matter so as to produce the fundamental metaphysical principles required for physics, Newtonian physics in particular, to be possible as a science. In the course of extending the scope of his transcendental philosophy to include the foundations of an empirical science, Kant provided a concrete illustration of the account of causality he had developed earlier and also both more specific causal principles and arguments in support of them.

In the Dynamics, for example, Kant develops a matter theory according to which matter, any spatial substance, can fill a determinate region of space through the exercise of its attractive and repulsive forces. Kant's theory is developed in explicit contrast with atomistic theories, such as Lambert's, according to which matter fills space by means of its absolute solidity, on the grounds that such an account does not truly explain how matter is extended. While the so-called 'balancing argument' that Kant uses to defend the postulation of specifically attractive and repulsive forces—positing only attractive forces would lead to everything collapsing at a point, while positing only repulsive forces would lead to infinite expansion—is unlikely to convince in the form in which it is presented, the way in which it illustrates mutual interaction turns out to be rather helpful. For his idea is that the repulsive force of one matter will resist the attempt of other matter to penetrate the space that it occupies, but because the other matter will resist any attempt to penetrate the space it occupies, it turns out that only through two substances exercising their repulsive forces jointly can the line of demarcation between the regions they occupy be determined. That is, any two bodies can occupy determinate (and neighbouring) regions in space only if they mutually interact through the joint exercise of their repulsive forces. Since a similar point holds for other features of matter, one can see that mutual interaction is actually more central to Kant's position in physics than is the simpler notion of causality discussed in the Second Analogy.

In the Mechanics, Kant advances three Laws of Mechanics that explain the communication of motion. The First Law asserts the conservation of matter throughout any changes in its motion, the Second Law is a version of the law of inertia, while the Third Law posits the equality of action and reaction in any changes of motion. While the latter two laws obviously involve causality in a direct way, the nature of Kant's arguments for these laws is quite striking. For what justifies these laws is the fact that only by means of them is experience of the communication of motion possible. In other words, just as the Analogies of Experience were transcendental arguments revealing the conditions on the possibility of a certain kind of temporal experience, the Laws of Mechanics are designed to establish conditions on the possibility of a particular kind of spatial experience (specifically, experience of how matter can communicate motion). As a result, we can see that Kant is in fact interested in a wide range of transcendental arguments; the more content that is contained in the notion of experience being considered, the more robust are the conditions of its possibility.[7] Accordingly, causality can play a role at a wide range of levels within Kant's transcendental philosophy.

4.2 The *Critique of the Power of Judgment:* Biology

While the *Metaphysical Foundations of Natural Science* focuses on efficient, mechanical causality, the *Critique of the Power of Judgment* (1790) devotes considerable attention to developing a detailed account of teleological or final causation. At a very high level of generality, Kant's strategy here is to acknowledge and characterize the fundamental differences between mechanical causality in physics and teleological causality in biology and even to privilege the former over the latter, but without denying that teleological causality has a legitimate place in our knowledge of the world. The central distinction that Kant uses to try to achieve this aim is that between constitutive and regulative principles, because mechanical causation is constitutive in the sense that, as we saw in the *Metaphysical Foundations*, it is required for the very possibility of experiencing matter, whereas the concept of a natural end that would be used in biology is not, even if it is helpful to us in unifying our experience of the world of living things.

FURTHER READING

As one might expect, given its systematic importance, a tremendous amount of secondary literature has been devoted to Kant's treatment of causality in the last two centuries, both in English and in other languages (especially German). Rather than attempt to give a comprehensive account of the various historical schools of interpretation (such as the numerous strands of Neo-Kantianism, including Cohen's, Cassirer's, Natorp's, Windelband's, etc.), I shall limit my overview to the narrow range of literature written in English during the past few decades.

The topic that has been discussed most frequently concerns the Second Analogy of Experience. Typically, this issue has been covered in book-length commentaries focusing on a larger context (most often that of the *Critique of Pure Reason* as a whole), with particularly notable contributions by Bird (1962), Strawson (1966), Melnick (1973), Beck (1978), Allison (1982), Guyer (1987), Longuenesse (1998), and van Cleve (1999). There are several significant articles as well, some of which can be found in a volume edited by Harper and Meerbote (1984). Friedman's (1992*a*) stimulating article on the topic should also be noted.

A second issue that has become more prominent recently is the metaphysics of causality. Langton's (1998) discussion of Kant's claim that we cannot know things in themselves rests on metaphysical issues about causality, and Watkins's (2005) book-length treatment of causality in Kant defends an interpretation emphasizing Kant's model of causality and its attendant metaphysics.

There have been a number of important discussions of causality in the philosophy of science. For the context of physics, Buchdahl (1969) and Friedman (1992*b*) are both classic treatments of the issue. Kant's views on teleology have recently been discussed in McLaughlin (1990), and in articles by Ginsborg and Guyer (e.g. in a volume edited by Watkins (2001)).

Finally, the relationship between Kant's views on causality and freedom and the way it is supposed to involve Transcendental Idealism have been discussed at length in innumerable

places. One prominent anti-metaphysical line of interpretation has been ably defended by Allison (1990) and developed in different ways by Korsgaard (1996) and Hudson (1994), while a more metaphysical approach has been articulated by Wood (1984) and developed in greater detail by Watkins (2005).

References

ALLISON, HENRY (1982). *Kant's Transcendental Idealism: An Interpretation and Defense* . New Haven: Yale University Press.

—— (1990). *Kant's Theory of Freedom.* New York: Cambridge University Press.

ARMSTRONG, DAVID (1983). *What is a Law of Nature?* New York: Cambridge University Press.

BECK, LEWIS WHITE (1978). *Essays on Kant and Hume,* ed. L. W. Beck. New Haven: Yale University Press.

BIRD, GRAHAM (1962). *Kant's Theory of Knowledge.* New York: Humanities Press.

BUCHDAHL, GERD (1969). *Metaphysics and the Philosophy of Science.* Oxford: Basil Blackwell.

FRIEDMAN, MICHAEL (1992*a*). 'Causal Laws and the Foundations of Natural Science', in P. Guyer (ed.), *The Cambridge Companion to Kant.* New York: Cambridge University Press, 161–99.

—— (1992*b*). *Kant and the Exact Sciences.* Cambridge, Mass.: Harvard University Press.

GINSBORG, HANNAH (2001). 'Kant on Understanding Organisms as Natural Purposes', in E. Watkins (ed.), *Kant and the Sciences.* New York: Oxford University Press, 231–58.

GUYER, PAUL (1987). *Kant and the Claims of Knowledge.* New York: Cambridge University Press.

—— (2001). 'Organisms and the Unity of Science', in E. Watkins (ed.), *Kant and the Sciences.* New York: Oxford University Press, 259–81.

HARPER, WILLIAM, and MEERBOTE, RALF (1984). *Kant on Causality, Freedom, and Objectivity.* Minneapolis: University of Minnesota Press.

HUDSON, HUD (1994). *Kant's Compatibilism.* Ithaca: Cornell University Press.

KORSGAARD, CHRISTINE (1996). *Creating the Kingdom of Ends.* New York: Cambridge University Press.

LANGTON, RAE (1998). *Kantian Humility.* New York: Oxford University Press.

LONGUENESSE, BEATRICE (1998). *Kant and the Capacity to Judge.* Princeton: Princeton University Press.

MCLAUGHLIN, PETER (1990). *Kant's Critique of Teleology in Biological Explanation* . Lewiston: Mellon.

MELNICK, ARTHUR (1973). *Kant's Analogies of Experience* . Chicago: University of Chicago Press.

STRAWSON, PETER (1966). *The Bounds of Sense.* New York: Methuen.

TOOLEY, MICHAEL (1977). 'The Nature of Laws', *Canadian Journal of Philosophy* 7: 667–98.

VAN CLEVE, JAMES (1999). *Problems from Kant.* New York: Oxford University Press.

WATKINS, ERIC (ed.) (2001). *Kant and the Sciences.* New York: Oxford.

—— (2005). *Kant and the Metaphysics of Causality*. New York: Cambridge University Press.

Wood, Allen (1984). 'Kant's Compatibilism', in A. Wood (ed.), *Self and Nature in Kant's Philosophy*. Ithaca: Cornell University Press, 73–101.

CHAPTER 6
THE LOGICAL EMPIRICISTS

MICHAEL STÖLTZNER

CAUSATION was a central theme for the movement of Logical Empiricism (hence-forth LE); during its classical European phase—the 1920s and 1930s—and beyond. It would not become one of LE's alleged 'dogmas', unlike verificationism and the analytic–synthetic distinction. Rather, the topic of causation paradigmatically exhibits two important features of LE. First, the movement was intimately connected to the scientific developments of the day; its representatives tried to accommodate their analyses to those developments rather than insist on an unassailable philosophical outlook come what may. Those Logical Empiricists who wrote the most on causality all had a physics background: Philipp Frank (1884–1966), Moritz Schlick (1882–1936), and Hans Reichenbach (1891–1953). They were constantly interacting with key members of the German-speaking physics community. Second, their joint allegiance to scientific empiricism and modern logic, and the common agenda to replace traditional metaphysics by a scientific world conception, cannot conceal the fact that the members of LE stemmed from different intellectual backgrounds and pursued, the manifold cross-references notwithstanding, original trains of thought. Hence they reacted in different ways to the scientific revolutions that occurred during the heyday of LE, quantum theory foremost.

Let me first list the elements that Frank, Schlick, and Reichenbach had in common. (1) The object of their studies of causality was the basic equations of physical science, not everyday or ordinary-language intuitions about causation. The general principle (or law) of causality, in first approximation, stated that any scientific fact could be described in such a form. (2) All three sought to make their theory of causation applicable to scientific progress. As Frank's comrade-in-arms Richard von Mises (1930: 146) put it, the law of causality 'is *changeable*, and it will *subordinate itself to the demands of physics*'. (3) All three were convinced that statistical physics could only be assessed on the basis of objective probabilities. Accordingly, there existed a close connection between their views on causality and probability. (4) In the 1930s, they criticized those interpretations according to which quantum mechanics had opened the door for a return of traditional 'metaphysical' ideas, among them vital factors and the freedom of the will. In doing so, they treated both relativity theory and quantum mechanics as the joint pillars of modern science, although they were well aware that the two theories had substantially different causal features.

The following topics, however, remained controversial: (1) Frank and, from the mid 1920s on, Reichenbach adopted the relative frequency interpretation of probability and allowed for the basic laws of nature being statistical. Schlick's preferred interpretation of probability, by

contrast, blocked any talk of probabilistic causality. (2) To both Frank and Schlick, a scientific theory, be it statistical or deterministic, represented a set of symbols and relations that had to be coordinated to experiences. To Reichenbach, however, statistical theories enjoyed a unique status in virtue of the statistical character of any scientific judgement. Reichenbach also stood against Frank and Schlick in anchoring the connection between causality and the direction of time at the microlevel. (3) While Frank intended to find an empirical meaning of the general principle of causality and uncovered its mainly pragmatic nature, Schlick and Reichenbach held that causal laws could not be applied without presuppositions of a non-empirical nature.

Only some of the ideas discussed in the present chapter influenced the debates after 1945. Among them are Reichenbach's causal forks and the common cause principle, and his ideas on the direction of time, all of which received their mature versions only in his *Philosophic Foundations of Quantum Mechanics* (1944) and *The Direction of Time* (1956). The contributions of Frank and Schlick, in contrast to the broad discussion they received in the European period, were now relegated to the footnotes and cited in an emblematic fashion. Moreover, the focus of philosophers of science had meanwhile shifted from causality and determinism in atomic physics to general issues of causal versus statistical explanation and inductive inference (see Salmon 1989). When interest in the philosophy of quantum mechanics returned in the 1970s, the discussions focused on the issue of realism, and LE was typically seen as a positivistic justification of the Copenhagen Interpretation that dogmatically forbade any search for a causal theory of atomic phenomena.

1. PHILOSOPHICAL INFLUENCES

Among the various philosophical influences on the members of LE, three are of special importance in matters of causality: Ernst Mach's reinterpretation of Hume, Kant's categorical conception as understood by the neo-Kantians, and the conventionalism advocated by Pierre Duhem and Henri Poincaré.

Already in his early work, Mach reinterpreted Hume's empiricist notion of causality—more precisely, the relation of 'constant conjunction'—as a mathematical function between a complex of determining conditions (or elements). Mach's notion of causality was not only holist and empiricist; he also rejected any counterfactual reasoning. 'There is no cause nor effect in nature; nature has but an *individual* existence' (1989: 580). Machian causality, moreover, did not involve the temporal order of phenomena.

From a Kantian or neo-Kantian perspective the price of Mach's liberal conception was too high. Since any stable functional dependency qualified as causal, the Machian had to find independently the entities or facts standing in a causal relation so conceived, while the Kantian framework admitted as 'empirically real' only that which fell under the category of causality. Because of this additional freedom, Machian causality, to the neo-Kantians, was pretty close to tautology.

The Machian and the neo-Kantian lines of thought combined with two different interpretations of probability. Von Kries's (1886) *Spielraum* theory, for one, was initially preferable for the neo-Kantians because it incorporated objective chance into a deterministic world by distinguishing strictly causal nomological regularities from ontological regularities

that existed within the range (*Spielraum*) admitted by the former. In this way, the categorical status of the principle of causality remained intact. By contrast, the relative frequency interpretation developed by Gustav Theodor Fechner (1897) and, more rigorously, by von Mises (1919) was based on the notion of statistical collective and dealt with the distributions of statistical variables within such collectives. Whether certain variables formed a collective was a purely empirical matter.

In the youthful days of Frank, Schlick, and Reichenbach the philosophical import of the Second Law of Thermodynamics was heavily debated. Had Boltzmann already demonstrated —as Franz Serafin Exner (1909) proclaimed in Vienna long before the advent of quantum mechanics—that chance is the basis of all natural events because the strict laws physicists observe only emerge as the macroscopic limit of a very large number of random microscopic events? Or was this strategy to overcome the fundamental dualism between reversible and irreversible phenomena, between dynamical and merely statistical lawfulness, 'a fatal and shortsighted mistake'—as Max Planck (1914: 23) retorted—that sacrificed the very foundation of the scientific enterprise, the search for strictly valid laws of nature? (See Stöltzner 1999; 2003.)

French conventionalism had emerged outside the context of the atomism controversy, from the late nineteenth-century debates as to whether physical space was, as Kant had assumed, necessarily Euclidean. Poincaré and Duhem, in contrast, argued that the choice of a geometry was conventional, and that among the empirically admissible conventions we usually pick the simplest one. LE developed this idea into a general method following which the basic terms of a scientific theory were related to observations by way of 'conventions' or 'coordinative definitions'. This method permitted them to strictly hold apart the scientific theory, now understood as a part of logic, from the empirical observations and accordingly to deny any space for the synthetic a priori.

2. CAUSALITY DURING THE EUROPEAN PHASE OF LE

In the 'First Meeting on the Epistemology of the Exact Sciences' that was held in Prague in 1929 and in which the Vienna Circle first went public with its famous manifesto (Verein Ernst Mach 1929), 'probability and causality' was a major topic. The papers, and the extensive discussion that followed them and was printed in *Erkenntnis* (Zilsel et al. 1930), showed that there existed a great variety of opinions. One can basically identify three groups: (1) Friedrich Waismann (1930) supported a Kriesian theory as regards both physical probabilities and the logical probability of judgement. But according to Waismann the two concepts of probability were distinct, and the so-called application problem (cf. Zilsel 1916)—that is, the question as to why the logico-mathematical structure of probability calculus was valid for events in the empirical world—was meaningless. Herbert Feigl (1930) did not share this latter view, but still favoured the Kriesian approach over frequentism. All held, as did Schlick, that on the frequentist route one would arrive at correct statements only in 'the limit for infinitely many cases—which for the empirical world is naturally a senseless requirement' (Schlick 1931: 158/200–1).[1] In the discussion, Rudolf Carnap endorsed Waismann's view about logical probabilities, but did not enter into the issue of causality. In an early paper (Carnap 1924) that stands within a neo-Kantian framework and adopts Hans Vaihinger's fictionalism, he had

argued that the dimensionality of space follows from the principle of causality. (2) Von Mises (1930), in close agreement with Frank (1929), advocated his own frequentist approach. 'Probability' in everyday parlance—our subjective degree of certainty—was to his mind sharply distinct from its mathematical homonym. Thus there was no application problem either. (3) Reichenbach linked physical probability with the probability of inductive inference thus bestowing statistical physics with a specific epistemic status as compared to any other scientific theory.

Probability would remain a controversial topic within LE, until Carnap (1945) proposed a kind of compromise. Starting from the actual use of probability, he argued that there were two distinct senses: a logical probability relating to judgement and inductive inference, the chief representatives being John Maynard Keynes and Harold Jeffreys, and a physical probability based on relative frequencies, developed most completely by von Mises and Reichenbach. By 1930, however, subjectivist or propensity interpretations of probability were hardly considered attractive in the context of physical causality. Kolmogorov's (1933) axiomatic formulation was not available either.

In the subsequent conferences organized by the LE, causality became a particularly fertile ground for interaction with leading scientists. In 1930, Heisenberg (1931) presented a paper that was followed by a discussion and a short bibliography on causality and probability. And in 1934, causality entered through the criticisms against the physicist Pascual Jordan's claim that quantum mechanics suggested the existence of vitalistic factors in physics.

The 1936 Copenhagen congress was explicitly dedicated to 'The Problem of Causality—with Special Considerations of Physics and Biology'. The physics section opened with a talk by Niels Bohr (1937) who outlined how his concept of complementarity replaced causality in the atomic domain. In this way, he had just responded to Einstein's criticism, the famous EPR thought experiment (Einstein, Podolsky, and Rosen 1935; Bohr 1935). Frank (1937) and Schlick (1937) tried to develop an empiricist reading of Bohr's ideas that was immune to metaphysical interpretations. In the biology and the psychology sections, one finds among others contributions by J. B. S. Haldane, Nikolai Rashevsky, and Edward C. Tolman. None of the three can be considered an advocate of LE. But Neurath's strategy in planning these congresses was to establish alliances with kindred groups in other countries and to integrate leading scientists. In the congresses after 1936 there is little mention of causality, an interesting exception being the legal positivist Hans Kelsen's (1939) theory that the law of causality emerged from the principle of retribution.

Schlick supervised three Ph.D. theses dedicated to causality, by Herbert Feigl, Marcel Natkin, and Tscha Hung (Haller and Binder 1999). Feigl's thesis contains interesting ideas, for example about the application problem and the relationship between causality and uniqueness, but it was never published because the author considered it refuted by the progress in physics.

3. PHILIPP FRANK AND THE SEARCH FOR AN EMPIRICAL MEANING OF CAUSALITY

At the time he wrote his first philosophical paper, Frank had just obtained his Ph.D. His

punchline was to exchange Kant for Poincaré. 'The law of causality ... can neither be confirmed nor disproved by experience; not, however, because it is an *a priori* true necessity of thought, but because it is a purely conventional definition.' Frank discussed the law of causality in the following form that goes back to Hume: 'If, in the course of time, a state of the universe *A* is once followed by the state *B*, then whenever *A* occurs *B* will follow it' (Frank 1907: 444–5/63). In this form, the law was only applicable to the universe as a whole or portions of it that were 'large enough' for the accuracy required. But there was a more profound arbitrariness. 'If the law of causality is not valid according to one definition of the state, we redefine the state simply in such a way that the law is valid. If that is the case, however, the law, which appeared to be stating a fact, is transformed into a mere definition of the word "state"' (1907: 446–7/65).

In 1919, Frank applied a similar analysis to statistical mechanics. 'If one understands by state the sum of all *physically measurable* properties of the system, the law of causality has no validity. In the sense of molecular theory one must rather add to the description of the state also the positions and velocities of all molecules, by means of which the law of causality is saved, but its actual application becomes impossible' (Frank 1919: 727). In the phenomenon of Brownian motion, however, one can actually observe the spontaneous fluctuations emerging from the randomness on the microscopic level. This gave his 1907 argument a new twist, in so far as it was possible to establish an average law in the mesoscopic domain. Frank now characterized the Second Law of Thermodynamics, to wit, that the entropy of any closed system increases in the mean, as the most characteristic trait of natural processes and as a 'brazen law' (1919: 701) that originates in a game of chance.

In his 1932 book *The Causality and Its Limits,* Frank called his earlier position one-sided and cited von Mises's conception of statistical law (*Gesetzmäßigkeit*)[2] and quantum mechanics as motivations to modify his stand. Frank's overall strategy was to search for conditions under which a suitably refined general law of causality attained an empirical meaning. He arrived at a negative conclusion:

From our experiences no proof can be derived for or against the validity, or even probability, of the law of causality in nature, nor can we conclude anything about observable events from the validity of the law of causality.

On the other hand, our whole science, even our whole practical life is apparently based on the continual application of the law of causality. ... Each manipulation is accompanied by the expectation of definite results. ...

Both conceptions are correct and therefore cannot be in real opposition. The appearance of such a contradiction comes about because ... an old tradition has taught us to look for a sharply designed world of 'real' things behind the living, but vague, world of our experiences.   (Frank 1932: 286–7/238–9)

The guilty party was the correspondence theory of truth. As Schlick (1925*b*) had shown, the only meaningful criterion of truth was the uniqueness of coordination between the mathematical symbols of physical theory and our sense experiences:

[T]he systematic summary of a [theoretical] scheme and rules of coordination can form a causal or a noncausal theory. The latter is true when the mathematical magnitudes are coordinated not to individual experiences, but, as happens in modern wave mechanics, to a whole group of experiences, which results from a series of experiments made under certain conditions.   (Frank 1932: 333–4/279–80)

The advantage of the new quantum theory was to include measurement errors, the summary nature of the prediction of experiences, into the theoretical scheme itself. Such was made possible by Mach's liberal notion of causality, following which mass phenomena qualified as basic entities on a par with electric charges and mass points. Frank was also open to future modifications of quantum mechanics, that is, to the possibility 'that we shall perhaps some day find a set of quantities with the help of which it will be possible to [causally] describe the behavior of these particles in greater detail than by means of the wave function, the frequencies' (Frank 1929: 992–3/123).

'[O]bviously the general law of causality was not a great discovery. Only special causal laws were, for example the discovery by Galileo and Newton that all motions can be predicted from the positions and velocities at a moment in time' (Frank 1932: 328/274–5). This ideal was expressed in Laplace's demon, 'probably the most incisive and definite [formulation the law of causality] has ever received' (ibid. 60/44). But in order to cash out pre*determination* into actual pre*diction*, Laplace's demon, when dealing with Newton's laws of motion, 'must know all initial positions of all mass-points of the world; he must know the forms of all [force] functions X, Y, Z for all masses, and finally he must be able, from knowledge of the initial conditions and of the functions X, Y, Z, to calculate the positions at any time whatsoever' (ibid. 64/48).

But already some theories of classical physics, among them hydrodynamics and elasticity theory, fall short of this ideal. 'For the state of the system is described by magnitudes that result from forming averages of positions and velocities, for example density or the shape of the surface' (ibid. 70/53). And Frank mentioned that von Mises (1922), accordingly, had advocated a genuinely statistical approach to hydrodynamics to overcome this 'crisis of mechanics'. Frank also discussed systems—today called 'chaotic'—which exhibit a sensitive dependence on initial conditions. 'In this sense we can say that the world of mechanical laws, if we want to pursue it into its finest details, has "gaps like a sieve"' (Frank 1932: 103/81).

Concerning the classical problem of miracles, Frank devised an argumentative strategy that he also applied against the vitalist claim that animate nature could only be explained teleologically. Diagnosing a miracle is non-tautological only if it is turned into a positive statement about the psychological states of a superior intelligence. Vitalists, to be sure, intend to avoid such theological connotations and phrase teleological explanations by analogy to scientific explanations. But then these explanations yield either testable predictions, in which case they are not at all different from causal laws, or they are tautological and resemble a proto-scientific theory close to animism.

Ludwig von Bertalanffy's attempts 'to formulate vitalism positivistically' (ibid. 147/117) faced the same dilemma. Representing a 'system' was no peculiar trait of living organisms.

'We cannot solve the most primitive mechanical problem, the path of a particle on which no forces have an effect, without knowing its velocity with reference to the whole galaxy' (ibid. 150/120). Also the fact that later states of a system enter in the description of its behaviour— if one applies an integral principle or in the case of transitions between atomic energy levels— did not indicate irreducibly teleological features. It was simply another way to arrive at a causal description because the law of causality did not specify a temporal order.

In the mid-1930s, Frank's main theme was to counter metaphysical misinterpretations of quantum mechanics. In his talk at the 1936 Copenhagen conference, he intended to give an empiricist reading of Bohr's complementarity. Rather than rehearsing Bohr's argument for the indispensability of classical concepts, he commenced from Otto Neurath's (1931) brand of physicalism and emphasized the special role of a purified everyday language for the unity of science. In this language used to describe gross-mechanical processes, 'all men are in harmony. ... Bohr has demonstrated ... that certain parts of the language of everyday life can nevertheless be retained for certain experimental arrangements in the field of atomic phenomena, although different parts are required for different experimental arrangements' (Frank 1937: 316/170). If these limits for the applicability of concepts are transgressed, metaphysical pseudoproblems emerge dealing with a 'real' world of particles having simultaneously definite positions and momenta that is, unfortunately, unknowable to us forever. As regards Bohr's extension of complementarity to the biological domain, Frank remained sceptical.

Frank's 1957 introductory *Philosophy of Science* in large part followed the reasoning just outlined. He now explicitly distinguished between a Humean formulation of the principle of causality based on the recurrence of states and a Kantian one based on the existence of laws, the former being much closer to tautology. And he also distinguished, more clearly than in 1932, between causal laws, among them Schrödinger's equation, and statistical laws. These were not mutually irreconcilable. 'If we speak in terms of observable phenomena, all laws are statistical' (Frank 1957: 296). But some have exactly causal laws as a purely mathematical limit, while others do not. In the case of Schrödinger's equation, 'we cannot predict single point-events at a definite location in space ... because the operational definition of Ψ does not link its value to single point-events, but to a statistical average computed from a great number of point-events' (ibid., 346). Frank endorsed Carnap's (1945) distinction between two probabilities as a compromise and emphasized that either concept required operational definitions in the sense of P. W. Bridgman (1927) to become applicable. Frank thus, in the end, no longer followed von Mises in deeming meaningless the application problem and all talk about the (subjective) probability of single events.

4. From Kant to Verificationism: Moritz Schlick's Two Theories of Causality

By 1920, Schlick was the leading German philosopher of relativity theory. This perspective dominated his first paper on causality, in which statistical mechanics was mentioned only in passing. Schlick started from a Kantian conception, holding that the principle of causality is 'a general expression of the fact *that* everything which happens in nature is subject without exception to valid laws' (Schlick 1920: 461/295). These laws are, as a purely conceptual structure, atemporal and discovered by man. Mathematically, they

correspond to sufficiently simple differential equations and their boundary conditions. This scheme is only well defined if action-at-a-distance does not occur in nature, which represents a fact about nature. But in order to arrive at causal laws other than by blind guesswork, one had to assume some a priori principles of uniformity and universality:

[I]f like cases are to be able to exist in the course of nature, some principle of separation must be presupposed, which sees to it that occurrences can be *alike* without being *identical*[:] ... spatial coexistence and temporal succession. ... These principles of separateness, which constitute the presupposition of the concept of lawfulness and causality, have rightly been called *forms*, following Kant's terminology.   (Schlick 1920: 467–8/307–8)

Space and time coordinates, accordingly, must not enter explicitly into the mathematical expression of the laws because otherwise there would exist no verifiable difference 'between a universe confounded by chance and one thrown into confusion by causality' (ibid. 465/303). And it would also contradict general relativity, or more precisely the principle of general co-variance.

'The causal determinacy of the world extends only in *one* dimension, and this we call the direction of time. Once it is chosen, what lies in the other three dimensions has to be seen as simply *contingent*' (ibid. 474/319). In what was left nomologically contingent, there could exist ontological regularities. Among Schlick's examples are the identity of all electrons in the world and the hypothesis of molecular disorder that implements the unidirectionality of time based on the second law of thermodynamics. That statistical regularity was a specific type of ontological regularity had been the starting point of von Kries's theory of probability.

In a 1925 textbook entry, Schlick rehearsed this theory of causality. But he did not conceal his lingering doubts as to whether atomic transitions could still be understood as ontological regularities. If the newest results of quantum theory were true, '[t]he world, in the last resort, would be handed over to chance' (Schlick 1925*a*: 460/60). Schlick did not welcome this development. '[O]nly in the utmost case of emergency will the scientist or philosopher decide to postulate purely statistical micro-laws. ... The principle of causality would be abandoned ... and hence the possibility of exhaustive knowledge would have to be renounced' (ibid. 461/61).

In 1926, the case of emergency did occur. Schlick, after five years of virtual silence on causality, revoked his earlier theory and called for a fresh start. '[E]*very* ordering of events in the temporal direction, of whatever kind it may be, is to be viewed as a causal relation. Only complete chaos, an utter lawlessness, could be described as non-causal occurrence, as pure chance' (Schlick 1931: 146/179). Also space and time coordinates may enter explicitly into a causal regularity. Schlick now shifted emphasis from how a law is established to how it is tested. If we have found a formula describing our observations, we investigate whether it 'also presents correctly those observations which we have *not yet used* in obtaining it. ... In other words, the true criterion of lawfulness [*Gesetzmäßigkeit*], the essential mark of causality, is the *fulfilment of predictions*'. In this, 'past and future data have entirely equal rights'. Schlick considered his new criterion as necessary and sufficient: 'only by means of it does reality speak to us; the establishing of laws and formulae is purely the work of man' (ibid. 149–

50/185; 150/186–7; 150/185).

However, 'confirmation of a prediction never ultimately *proves* the presence of causality, but always makes it probable, merely. ... From this we gather that a causal claim by no means has the logical character of an *assertion* [*Aussage*], for a genuine assertion must ultimately allow of verification.' Rather, it represents a 'prescription for the making of assertions' in the sense of Wittgenstein. For this reason, the principle of causality cannot be empirically true or false, but represents a directive 'to seek regularity [*Regelmäßigkeit*] and to describe events by means of laws' (ibid. 150/187; 151/188; 155/196).

There exist no general criteria for a law being causal. We can look only at the single act of verification. '[F]ulfilment of a prophecy, is an ultimate, incapable of further analysis' (ibid. 151/188). Surprisingly, this move paved the way for a return, in pragmatic clothes, of those criteria previously dismissed for the definition of causality. It is 'probable' that a formula fulfilling the uniformity condition and the criterion of simplicity 'really expresses a law, an order that actually exists, and hence that it will be *confirmed*. ... The term "probability", that we use here, means something quite different ... from the concept that ... occurs in statistical physics'. Thus, simplicity '*de facto* coincides with the true criterion [of causality], that of *confirmation*. It obviously represents, in fact, the special prescription, effective in our world, whereby the general directive of the principle of causality, to seek regularity, is supplemented' (ibid. 151/187–8; 156/197–8).

Turning to quantum mechanics, Schlick emphasized that Heisenberg's uncertainty relation involved an indeterminacy of prediction.

The new contribution made by present physics to the causality problem does not consist in contesting the validity of the principle of causality as such ... nor in the recognition of a purely probabilistic validity of natural laws having replaced belief in their absolute validity. ... The novelty, rather, consists in the discovery, never previously anticipated, that a limit of principle is set to the exactness of prediction by the laws of nature themselves. (ibid. 153/191)

Nonetheless, Schlick upheld the distinction between strict law and sheer randomness, which replaced the earlier one between nomological and ontological regularities and allowed him to maintain the Kriesian definition of probability. A statistical law that is confirmed in, say, 99 per cent of cases is

the resultant of two components, in that the imperfect or statistical causality is dissected into a strict regularity [*Gesetzmäßigkeit*] and an element of pure chance, which overlap. ... In the kinetic theory of gases, the laws whereby each individual particle moves are assumed to be totally rigorous; but the distribution of individual particles and their states is presumed at any given moment to be entirely 'lawless'. ... [In quantum mechanics], the description of processes is likewise split into two parts: into the lawful propagation of Ψ-waves, and into the occurrence of a particle or quantum, which is absolutely random, within the limits of the 'probability' determined by the Ψ-value at the point in question. (ibid. 157/198–9)

Accepting statistical laws without performing this spilt, as Reichenbach (1925) had done,

corresponded, in Schlick's view, to a mere definition of 'chance'. Schlick's reluctance to accept statistical laws puzzled all the physicists he had invited to comment upon his paper. Heisenberg, for one, wondered about the meaning of 'absolutely random, within the limits of probability' and contrasted the statistical quantum laws to Einstein's demand for a strict causality on the atomic level. But also Einstein did not believe that 'statistical law' was a contradictory concept.[3] In the face of this criticism, Schlick abandoned the idea of splitting statistical causality into law and randomness, but he maintained the Kriesian concept of probability.

Still, in his 1936 Copenhagen paper, Schlick held that the 'uncertainty relations establish for the experimental results ... a quite specific range (*Spielraum*), having objective significance' (Schlick 1937: 320/485). But the paper took a new tack in so far as Schlick, in line with Frank, focused more strongly on language. In 1931 he still accepted sharp particle trajectories 'as a *façon de parler*' (Schlick 1931: 151–2/189). But now he tightened the linguistic strictures and held that the uncertainty relation 'does not mean that a perfect insight into existing connections is easily barred to us; it means that certain connections simply do not exist' (1937: 319/483–4).

The main purpose of this finality claim was to block metaphysical misinterpretations of quantum mechanics, among them talk about 'measurement' if it involved the 'psycho-physical relation' (ibid. 318/483). And Schlick also rejected tampering with the *tertium non datur*. Undetermined properties—as introduced by Reichenbach's probabilistic topology (see below) and quantum logic—to his mind endangered a core principle of consistent empiricism, 'that nothing in the world is *intrinsically* unknowable. There are many questions, to be sure, which for practical or technical reasons will never be answered, but a question is intrinsically insoluble only in the one case where it is no question at all' (ibid. 326/489–90) because its verification is excluded on purely logical grounds. Schlick was convinced that Heisenberg's uncertainty relation and Einstein's definition of simultaneity implied such a logical impossibility.

5. Causality, Probability, and Inductive Inference: Hans Reichenbach

As early as his Ph.D. thesis of 1915, Reichenbach developed two ideas that would remain central to his philosophy—while their epistemological status underwent significant changes. First, the principle of causality, to become at all applicable to the description of physical phenomena, must be supplemented with a second principle, then called the principle of lawful distribution or the principle of the continuous probability function. Second, there existed no fundamental difference between the theory of error presupposed by any measuring science and the probabilistic theories of physics. This implied that dynamical (strictly causal) and statistical laws were understood as lawful within the same conceptual framework without adopting an empiricist indeterminism Vienna-style.

In his early works, Reichenbach (1916; 1920*a*; 1920*b*) considered both the principle of causality and the principle of lawful distribution as synthetic a priori. Causality guaranteed that there existed a functional form of every physical judgement, while the principle of lawful distribution guaranteed the existence of definite values of the physical magnitudes figuring in

these equations because it implied that the disturbing factors, the 'irrational remainder of the determinants' (1920*b*: 148/315) not expressed in any additional laws, remained small enough. By doing so the probability function 'connects the events laterally' (1920*b*: 152/324). Reichenbach's second principle, accordingly, imposed an order on what Schlick (1920), following von Kries, had termed the ontological regularities.

Notice that in contrast to his philosophy of space and time, Reichenbach did not historically relativize these synthetic a priori principles. Still, the departure from Kant's original doctrine was substantial because in virtue of the second principle all physical laws, at least on the empirical level, became merely probable. On the other hand, genuinely probabilistic laws could thus be treated as objective laws of nature rather than 'escape routes sought out by the physicist when he lacks a more precise knowledge of the connections involved' (Reichenbach 1920*b*: 153/326). In this vein, Reichenbach (1920*a*) criticized Laplace's principle of insufficient reason (or 'principle of indifference') because it was unable to provide any positive reason for equiprobability. The equiprobability hypothesis was not even required to justify the principle of lawful distribution. Taking up Poincaré's analysis of games of chance, Reichenbach argued that one only needs the hypothesis that the values of a probability function $f(\Omega)$ are equally probable if the values of Ω are infinitely close. Or in other words, $f(\Omega)$ must be continuous while its special form is irrelevant. Since this hypothesis transcends all finite experiences, it represents a synthetic a priori principle that guarantees the applicability of probability calculus.

In a paper written in 1923 but published only in 1933 Reichenbach had partly changed his mind. Although he considered the a priori conception of causality as irrefutable, because one could still claim the existence of causal laws that have not been found to date, the principle of causality was not positively required for the existence of natural laws. The application of the other principle, which was now called probabilistic (or inductive) inference [*Wahrscheinlichkeitsschluß*], however required that causality was not a priori excluded.

At the beginning of the paper, Reichenbach called causality a complex of principles and provided a non-exhaustive list of its elements. It contained Schlick's (1920) assertion that spacetime coordinates must not figure in the laws themselves, the principle of action by contact, and the temporal order of events. But all three were only partial claims that supplemented the more general inductive principle of causality. This 'says that by means of a functional relationship unobserved events can be predicted from observed ones, no matter whether the observed events lie in the future, or in the past, or happen at different space points simultaneously with the act of observation' (Reichenbach 1933: 34/347). As he made clear in a rather similar list in his entry for the *Handbuch der Physik* that was written in the mid 1920s, causality is not exhausted by the concept of a function, as Mach had held, because it represents 'a functional connection of a very specific character' (Reichenbach 1929: 59/193). Laplacian determinism, on the other hand, represented an extrapolation beyond the implication from causes to effects and neglected the probabilistic character of knowledge. Reichenbach, having abandoned the synthetic a priori, in the mid 1920s positioned himself between the Kantian and the Machian tradition: the principle of causality could be empirically false but the principle of probabilistic or inductive inference remained a condition for the possibility of scientific knowledge.

Reichenbach's inductive principle of causality operated as follows: Starting from a

presumed law $F_r(p_1, ..., p_r)$ we find further relevant causes $p_{r+1}, ..., p_{r+s}$ that lead to a modified function F'_{r+s} $(p_1, ..., p_r, p_{r+1}, ..., p_{r+s})$. This new governing function is the simplest function that, without being ad hoc, approximates the additional parameters in the least squares. Iterating this procedure with new classes of observed points M'', M''', ... we obtain either the infinite governing sequence (I) $F_r, F'_{r+s}, F'_{r+s}, F'_{r+s}, ...$ or (II) $F_r, F'_{r+s}, F''_{r+s+t}, ..., F^{(i)}_{r+s+t+ ... +w}, ...$ In case (I) we have found a causal law, whereas in case (II) the connection between the observations is random. Both cases 'characterize an objective state of affairs' (Reichenbach 1933: 43/354), a conclusion for which the requirement of inductive simplicity is crucial. Otherwise, (I) could trivially be obtained by an arbitrarily complex function. Inductive simplicity also implied that the intermediate values between two observed values were described by F'_{r+s}. This assumption of smoothness shows that inductive simplicity has taken the place of the principle of the continuous probability function. To sum up: 'Either no continuous causal laws exist or they can be obtained by the requirement of simplicity' (ibid. 51/361).

Other than descriptive simplicity, which guided the choice by convention of a geometry in relativity theory, inductive simplicity represented a hypothesis about nature. But, the series (I) and (II) are infinite while further observations yield only finitely many data points. Thus we know, in a given case, only with probability whether causality holds or not. 'If causality holds in other cases, the probability that causality holds in the specific case under consideration merely *increases*' (ibid. 60/367). But it never actually reaches unity, that is, certainty. 'It is therefore not impossible that physics will some day be confronted by phenomena that compel it to abandon causality.' Mentioning quantum theory, Reichenbach concluded that '[i]n principle, it is possible to determine on the basis of experience whether causality holds' (ibid. 63/370).

Little wonder that, when finally publishing the paper with a decade of delay, Reichenbach proudly announced that meanwhile quantum mechanics had led to a breakthrough of his conception by providing a physical theory of type (II). Interestingly, in the handbook entry he had argued instead, more cautiously and in the same vein as Schlick (1931), that Heisenberg's uncertainty relations were 'an entirely new kind of restriction to our knowledge of nature, the existence of which was never before suspected' (Reichenbach 1929: 78/216).

Reichenbach's 1925 paper on the causal structure of the world was ambitious. Entirely dispensing with the hypothesis of strict causality, he proposed a conception based on 'the concept of probable determination alone'. This conception 'accomplishes everything that is achievable by physics and ... furthermore possesses the capacity to solve the problem of the difference between past and future, a problem to which the strict causal hypothesis has no solution'. Although he maintained his earlier convictions that physics rests upon both 'the principle of causal connection and the principle of probable distribution', and that one can in principle separate the causal connection between the determining factors and the probabilistic distribution of the remaining factors, he now considered the latter division as purely formal. It 'can be replaced by the single assumption that a connection of a probabilistic nature exists between cause and effect ...' This was precisely the inverse of Schlick's (1931) reasoning. By comparing all dependencies in the universe to a 'throw of a die', Reichenbach came pretty close to Exner's (1909) above-sketched reading of Boltzmann's legacy (Reichenbach 1925:

136/83; 133/81; 135/82; 138/84; 138/85).

By contrast with the Viennese empiricists, Reichenbach took a logical tack and replaced the causal connection of events by 'A implies B with probability', or $A \Rrightarrow B$, which he understood as a primitive concept. While logical implication ($\Rightarrow$) connects propositions, probability implication ($\rightarrow$) connects events. The most striking formal novelty was that $(A \Rrightarrow B) \rightarrow (A \Rrightarrow \neg B)$. One thus obtained a topology of probability implications, while the probability measure remained unspecified.

This topology, Reichenbach claimed, was sufficient to define a temporal order of events. 'If probability implication is valid in only one direction [i.e. $(B \Rrightarrow A) \wedge \neg(A \Rrightarrow B)$], then the antecedent [B] is the temporarily later event' (1925, 150/94). The main difference was that '[n]othing short of the totality of all causes is required for inferences into the future, but inference about the past can be made on the basis of a partial action [of causes]' (1925: 151/96). The future was thus objectively undetermined, a fact that led to a dispute between Schlick (1931) and Reichenbach (1931). In the 1925 paper, Reichenbach provided a detailed analysis of various inferential scenarios between three or more causes in the form of causal forks. This approach in its mature, and more rigorous, form outlined in the posthumous *The Direction of Time* (1956) became pretty influential on the debates about causality in the 1960s and 1970s.

In the handbook entry, Reichenbach also discussed the relationship between causality and the special theory of relativity on the basis of his method of mark transmission (Reichenbach 1924). A mark represented a small variation in an event. If we attach a mark to the cause *A*, this mark will also be observable in the effect *B*, but not vice versa. This asymmetry is 'the distinctive characteristic of the causal relation ... [and] can, in turn, be used in defining the sequence of time'. Accordingly, the 'objective significance of time consists in its formulating the type of order of causal chains' (Reichenbach 1929: 53/186; 57/190). And, referring to his 1925 definition, he argued that the microscopic events in nature could be subjected to temporal order. Boltzmann's contention that irreversibility and the direction of time emerge only as statistical features at the macro-level while atomic collisions remain reversible as in Newtonian mechanics, was, to Reichenbach's mind, too closely connected to the false ideal of Laplacian determinism. And Reichenbach (ibid. 62/196) criticized Schlick's (1925*a*) claim 'that every indication of temporal direction must conform to the Boltzmann scheme'.

Reichenbach (1929: 29/155) also criticized Schlick's characterization of truth as unique coordination. For this 'does not offer any means whereby the truth of a given physical proposition can be tested'. The only way to do so was by means of probability implications. 'We will no longer be able to speak strictly of the truth of a proposition, but only of its degree of probability' (ibid. 29/155). As shown above, unique coordination represented the basis on which Frank argued that there was no categorical difference between dynamical and statistical laws. To him and to von Mises, Newtonian mechanics, relativistic geometry, and (classical or quantum) probabilistic physics thus stood on a par. Reichenbach vehemently opposed this comparison: 'In the case of geometry, it is true, one is allowed to separate the problem of coordination from the mathematical theory because the problem of coordination does not contain any *geometrical* concept, but in the theory of probability the concept constituted by this theory enters into the problem of coordination' (Zilsel et al. 1930: 275). In his paper, Reichenbach (1930) advocated the relative frequency interpretation more explicitly than

before, and conjectured that every assertion of probability could be translated as an assertion of frequency. And in 1931 he openly criticized Schlick for remaining committed to the *Spielraum* interpretation. The association between finite observations and an infinite collective in the frequentist account, to Reichenbach's mind, was based on probability (or inductive) inference. But here von Mises, rightly, retorted that this association 'was not translatable into a frequency statement' (Zilsel et al. 1930: 282).

To avoid this criticism, Reichenbach (1930: 170) shifted the problem to the most basic level. 'Probability logic cannot be squeezed into the Procrustes bed of strict logic' which leads to the 'catastrophe of undecidability' about whether a law of nature is actually confirmed. Probability logic embraces strict logic as a limit in the same vein as truth appears as the limit of high probability. Probability logic can only be justified by 'the fact that we cannot think differently'. Moreover: 'The statement that probability laws do not hold is equivalent to predicting that, in repeated sequences, the regularity implied by the principle of induction does not hold—and this statement is empirically meaningful only if it can be decided inductively, that is, if the principle of induction holds. The statement that probability laws do not hold is thus self-contradictory and makes no sense' (ibid. 187/343; readjusted to original).

Since Reichenbach did not presuppose strict logic to hold, this contradiction did not amount to an indirect proof of the principle of induction. Rather, it finally dissolved Hume's problem, as Reichenbach proudly announced. To my mind, Reichenbach in effect treated induction—in the same vein as the principle of lawful distribution a decade before—as a condition for the possibility of experience, the only difference from a Kantian category being that no transcendental argument was available to justify it. Since he granted on the other hand that the principle of causality could be empirically inadequate, it appears that the two principles had changed rank. While initially the second principle—be it of lawful distribution or probabilistic inference—had only represented an indispensable complement to causality, it had now assumed the lead.

Given his repeated claims to have presaged important epistemological characteristics of quantum mechanics, it is quite surprising that in those years Reichenbach did not embark on a more detailed discussion of it and only criticized two interpretative claims of Heisenberg's. First, the 'positivistic' maxim to omit unobservable quantities from the theory 'must be correctly reformulated as the stipulation that *dispensable* quantities should be eliminated'. Yet this, so Reichenbach, was a simple consequence of probability inference. Second, Heisenberg's elucidation of the uncertainty relation as a disturbance effect, that is, that 'the *influence of the instruments of observation* cannot be ignored ... is not viable' (Reichenbach 1929: 78/215). For, he argued a year later, 'separation in object of observation and means of observation is an idealization that is to a certain extent fulfilled for certain macroscopic phenomena, but it cannot be regarded as a necessary presupposition of the exact sciences in the sense of the principle of causality' (1930: 180–1/338; transaltion readjusted to German original). The crucial point was rather that one could not push the probability to predict certain combinations of parameters arbitrarily close to unity.

Having emigrated to the United States, however, Reichenbach published extensively on the subject; and his 1944 book *Philosophic Foundations of Quantum Mechanics* inspired the subsequent debates. Although he still considered the 'quantum mechanical criticism of causality ... as the logical continuation of a line of development which began with the

introduction of statistical laws into physics' (1944: 3), he no longer embraced them within a common programme and emphasized the peculiarities of the quantum world. The central claim of the book was that causal anomalies were unavoidable if one insisted that interphenomena—that is, the states between the phenomena actually observed—possess definite values. Reichenbach's definition of a normal system was based on the idea that neither the laws of nature nor the states depend upon their being observed—while in 1930/1 he rejected precisely this kind of separation between object system and measurement apparatus.

The second major innovation of the book was Reichenbach's three-valued semantics for quantum mechanical statements. He was dissatisfied with the Copenhagen criterion for physically meaningful statements because this restriction was of a meta-linguistic kind, and physics could not get by without any description of interphenomena. The first dissatisfaction shows once again Reichenbach's distance from Schlick (1937) and Frank (1937).

In his posthumously published *The Direction of Time* (1956), Reichenbach modified his causal theory of time because it became clear to him that the mark method was not free of temporal concepts. At the very end of the book, Reich-enbach was worrying whether his idea of basing time on the microscopic order was faulted by R. P. Feynman's contention that a positron corresponded to an electron going backwards in time. In this way, a definite causal chain would exist merely locally and causal loops could not be excluded. (See Ryckman 2007 for a more detailed discussion of Reichenbach's late conceptions of causality.)

FURTHER READING

Virtually all the original texts used for this article are available in English translation in the Vienna Circle Library (now with Springer). Perhaps the best reading would be Frank's (1932) book, Schlick's (1931) second theory of causality, and Reichenbach's (1931) survey paper and his posthumous book (1956). So far, scholars have mainly concentrated on the contributions of LE to the philosophy of space and time. Ryckman (2007) provides a short overview of these developments and also discusses some aspects of the causality debate. Stadler (2001) provides a historical introduction into the Vienna Circle and supplies a wealth of biographical and bibliographical material on LE. Thematic analysis can be found in the *Cambridge Companion to the Vienna Circle* (Richardson and Uebel 2007).

REFERENCES

BOHR, N. (1935). 'Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?', *Physical Review* 48: 696–702.
—— (1937). 'Kausalität und Komplementarität', *Erkenntnis* 6: 293-303; English version in *Philosophy of Science* 4 (1937): 289–98.
BRIDGMAN, P. W. (1927). *Logic of Modern Physics*. New York: Macmillan.
CARNAP, R. (1924). 'Dreidimensionalität des Raumes und Kausalität. Eine Untersuchung über den logischen Zusammenhang zweier Fiktionen', *Annalen der Philosophie und philosophischen Kritik* 4: 105–30.
—— (1945). 'Two Concepts of Probability', *Philosophy and Phenomenological Research* 5: 513–32.

EINSTEIN, A., PODOLSKY, B., and ROSEN, N. (1935). 'Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?', *Physical Review* 47: 777–80.

EXNER, F. S. (1909). *Über Gesetze in Naturwissenschaft und Humanistik*. Vienna: Hölder.

FECHNER, G. T. (1897). *Kollektivmaßlehre*, ed. G. F. Lipps. Leipzig: W. Engelmann.

FEIGL, H. (1930). 'Wahrscheinlichkeit und Erfahrung', *Erkenntnis* 1: 249–59.

FRANK, P. (1907). 'Kausalgesetz und Erfahrung', *Ostwald's Annalen der Naturphilosophie* 6: 443–50; English trans. in Frank 1961: 62–8.

—— (1919). 'Die statistische Betrachtungsweise in der Physik', *Die Naturwissenschaften* 7: 701–5, 723–9.

—— (1929). 'Was bedeuten die gegenwärtigen physikalischen Theorien für die allgemeine Erkenntnislehre', *Die Naturwissenschaften* 17: 971–7, 987–94; also in *Erkenntnis* 1: 126–57. English trans. 'Physical Theories of the Twentieth Century and School Philosophy', in Frank 1961: 96–125.

—— ([1932] 1988). *Das Kausalgesetz und seine Grenzen*. Frankfurt am Main: Suhrkamp (1st pub. Vienna: Springer); English trans. Dordrecht: Kluwer, 1988.

—— (1937). 'Philosophische Deutungen und Mißdeutungen der Quantentheorie', *Erkenntnis* 6: 303–17; English trans. in Frank 1961: 158–70.

—— (1957). *Philosophy of Science: The Link Between Science and Philosophy*. Engleford Cliffs, NJ: Prentice Hall.

—— (1961). *Modern Science and Its Philosophy*. New York: Collier.

HALLER R., and BINDER, T. (eds.) (1999). *Zufall und Gesetz: Drei Dissertationen unter Schlick*. Amsterdam: Rodopi.

HEISENBERG, W. (1931). 'Kausalgesetz und Quantenmechanik', *Erkenntnis* 2: 172–82.

KELSEN, H. (1939). 'Die Entstehung des Kausalgesetzes aus dem Vergeltungsprinzip'. *Erkenntnis* 8: 69–130.

KOLMOGOROFF, A. N. (1933). *Grundbegriffe der Wahscheinlichkeitsrechnung*. Berlin: Springer.

MACH, E. (1989). *The Science of Mechanics: Account of Its Development*. La Salle: Open Court. (First published as *Die Mechanik in ihrer Entwickelung: historisch-kritisch dargestellt*. Leipzig: Brockhaus, 1883.)

NEURATH, O. (1931). 'Physicalism: The Philosophy of the Viennese Circle', *The Monist* 41: 618–23.

PLANCK, M. (1914). *Dynamische und statistische Gesetzmäßigkeit*. Leipzig: Barth.

REICHENBACH, H. (1916). *Der Begriff der Wahrscheinlichkeit für die mathematische Darstellung der Wirklichkeit*. Leipzig: Barth, offprint from *Zeitschrift für Philosophie und philosophische Kritik* (161: 210–39; 162: 9–112, 223–53). An author's report appeared in *Die Naturwissenschaften* 7 (1919): 482–3.

—— (1920*a*). 'Die physikalischen Voraussetzungen der Wahrscheinlichkeitsrechnung', *Die Naturwissenschaften* 8: 46–55; 'Nachtrag' p. 349; English trans. in Reichenbach 1978: ii. 293–311.

—— (1920*b*). 'Philosophische Kritik der Wahrscheinlichkeitsrechnung', *Die Naturwissenschaften* 8: 146–53; English trans. in Reichenbach 1978: ii. 312–27.

—— (1924). *Axiomatik der relativistischen Raum-Zeit-Lehre*. Brunswick: Vieweg.

—— (1925). 'Die Kausalstruktur der Welt und der Unterschied von Vergangenheit und Zukunft', *Sitzungsberichte der Bayerischen Akademie der Wissenschaften, mathematischnaturwissenschaftliche Abteilung,* 133–75; English trans. in Reichenbach 1978: ii. 81–119.

—— (1929). 'Ziele und Wege der physikalischen Erkenntnis', in *Handbuch der Physik,* iv. *Allgemeine Grundlagen der Physik.* Berlin: Springer, 1–80; English trans. in Reichenbach 1978: 120–225.

—— (1930). 'Kausalität und Wahrscheinlichkeit', *Erkenntnis* 1: 158–88; partially trans. in Reichenbach 1978: ii. 333–44.

—— (1931). 'Das Kausalproblem in der Physik', *Die Naturwissenschaften* 19: 713–22; English trans. in Reichenbach 1978, i. 326–42.

—— (1933). 'Die Kausalbehauptung und die Möglichkeit ihrer empirischen Nachprüfung', *Erkenntnis* 3: 32–64; English trans. in Reichenbach 1978: ii. 345–71.

—— (1944). *Philosophic Foundations of Quantum Mechanics.* Berkeley: University of California Press.

—— (1956). *The Direction of Time.* Berkeley: University of California Press.

—— (1978). *Selected Writings* 1909–1953. 2 vols. Dordrecht: Reidel.

RICHARDSON, A., and UEBEL, T. (eds.) (2007). *The Cambridge Companion to the Vienna Circle.* Cambridge: Cambridge University Press.

RYCKMAN, T. A. (2007). 'Logical Empiricism and the Philosophy of Physics', in Richardson and Uebel 2007: 193–227.

SALMON, W. C. (1989). *Four Decades of Scientific Explanation.* Minneapolis: University of Minnesota Press.

SCHLICK, M. (1920). 'Naturphilosophische Betrachtungen über das Kausalprinzip', *Die Naturwissenschaften* 8: 461–74; English trans. Schlick (1979: i. 295–321).

—— (1925*a*). *Naturphilosophie,* in Max Dessoir (ed.), *Lehrbuch der Philosophie: Die Philosophie in ihren Einzelgebieten.* Berlin: Ullstein, 397–492; English trans. in Schlick 1979: ii. 1–90.

—— (1925*b*). *Allgemeine Erkenntnislehre.* 2nd edn. Berlin: Springer.

—— (1931). 'Die Kausalität in der gegenwärtigen Physik', *Die Naturwissenschaften* 19: 145–62; English trans. in Schlick 1979: ii. 176–209.

—— (1937). 'Quantentheorie und Erkennbarkeit der Natur'; *Erkenntnis* 6: 317–26; English trans. in Schlick 1979: ii. 482–90.

—— (1979). *Philosophical Papers,* ed. Henk Mulder and Barbara F. B. van de Velde-Schlick. 2 vols. Dordrecht: Reidel.

STADLER, F. K. (2001). *The Vienna Circle. Studies in the Origins, Development, and Influence of Logical Empiricism.* Vienna: Springer.

STÖLTZNER, M. (1999). 'Vienna Indeterminism: Mach, Boltzmann, Exner', *Synthese* 119: 85–111.

—— (2003). 'Vienna Indeterminism II: From Exner to Frank and von Mises', in P. Parrini, W. Salmon, and M. Salmon (eds.), *Logical Empiricism. Historical and Contemporary Perspectives.* Pittsburgh: University of Pittsburgh Press, 194–229.

VEREIN ERNST MACH (1929). *Wissenschaftliche Weltauffassung. Der Wiener Kreis.* Vienna: A. Wolf. English trans. (without bibliography), *The Scientific Conception of*

the World: *The Vienna Circle*. Dordrecht: Reidel, 1973.

VON KRIES, J. (1886). *Prinzipien der Wahrscheinlichkeitsrechnung*. Freiburg i. B.: Mohr. 2nd edn. with a new foreword, 1927.

VON MISES, R. (1919). 'Fundamentalsätze der Wahrscheinlichkeitsrechnung', *Mathematische Zeitschrift* 5: 52–99, 100.

—— (1922). 'Über die gegenwärtige Krise der Mechanik', *Die Naturwissenschaften* 10: 25–9.

—— (1930). 'Über kausale und statistische Gesetzmäßigkeit in der Physik', *Die Naturwissenschaften* 18: 145–53; also in: *Erkenntnis* 1: 189–210.

WAISMANN, F. (1930). 'Logische Analyse des Wahrscheinlichkeitsbegriffs', *Erkenntnis* 1: 228–48.

ZILSEL, E. (1916). *Das Anwendungsproblem: Ein philosophischer Versuch über das Gesetz der großen Zahlen und die Induktion*. Leipzig: Barth.

—— et al. (1930). 'Diskussion über Wahrscheinlichkeit', *Erkenntnis* 1: 260–85.

PART II
STANDARD APPROACHES TO CAUSATION

CHAPTER 7
REGULARITY THEORIES

STATHIS PSILLOS

1. INTRODUCTION

David Hume has made available a view of causation as it is in the world that can be called the *Regularity View of Causation* (RVC). His famous first definition runs thus: 'We may define a *CAUSE* to be "An object precedent and contiguous to another, and where all the objects resembling the former are plac'd in like relations of precedency and contiguity to those objects, that resemble the latter" ([1739] 1978: 170).

More generally, we can present the kernel of RVC thus:

RVC
c causes *e* iff

  i. *c* is spatiotemporally contiguous to *e*;

 ii. *e* succeeds *c* in time; and

iii. all events of type *C* (i.e. events that are like *c*) are regularly followed by (or are constantly conjoined with) events of type *E* (i.e. events like *e*).

On RVC, the constitutive elements of causation are spatio-temporal contiguity, succession, and regularity (constant conjunction). Causation, that is, is built up from non-causal facts, more specifically two particular facts and one general. A corollary of RVC is that there is no extra element in causation which is of a fully distinct kind, like a necessary connection or a productive relation or what have you—something, moreover, that would explain or ground or underpin the regular association.

RVC has been espoused by many eminent philosophers and has been taken to be the official *Humean* view. I believe the Humean view was *Hume*'s view too, give or take a bit; for more on the reasons why, see Psillos (2002: ch. 1). However, there is strong opposition that has painted a picture of a new Hume: Hume was a causal realist, albeit a sceptical one. This interpretative line was inaugurated in the early 1970s—see John P. Wright (1973). (For an excellent recent discussion of Hume's theory of causation, advancing a projectivist interpretation of his views, see Beebee (2006*a*). See also [Chapter 4](#) of this volume.) In any case, as I shall argue in sect. 3.1, the first serious advocate of Humeanism was the Scottish philosopher Thomas Brown.

In this chapter I will articulate RVC with an eye to two things: first, its conceptual development; second, its basic commitments and implications for what causation is. I have chosen to present RVC in a way that respects its historical origins and unravels the steps of its articulation in the face of objections and criticism. It is important for the explication and defence of RVC to see it as a view of causation that emerged in a certain intellectual milieu. RVC has been developed as an attempt to remove efficiency from causation and hence, to view causation not as a productive relation but as a relation of dependence among discrete events. In particular, the thought that causation is regularity is meant to oppose metaphysical views of causation that posit powers or other kinds of entity that are supposed to enforce the regularities that there are in the world or to explain the alleged necessity that there is in causation.

Challenging the plausibility (and viability) of these metaphysically thicker accounts of causation has always been (and still is) part of the conceptual arsenal of RVC. By the same token, RVC is *not* fundamentally opposed to other theories that view causation as a relation of dependence and can certainly draw on their resources to develop a fully adequate account of causation. (For the distinction between causation as a relation of dependence and causation as a productive relation, see Psillos 2004 and 2007.) This means that the counterfactual theory of causation, or the probabilistic theory, are not, from the metaphysical point of view, rivals of RVC, though there are significant differences between them and they may well compete for which is the best account of causation among those that do not take it to be a productive relation.

RVC might well need supplementation to be an adequate theory of causation. But it is a central claim of the defenders of RVC that this won't remove the basic metaphysical credo of RVC, namely, that there is no necessity in nature and that, ultimately, causation depends, in some sense or other, on regularities. RVC should not be seen as a theory of the meaning of causal statements. Rather, it should be seen as a theory of what causation is in the world (better: a theory about the worldly component of causation)—but a theory whose metaphysical contours are constrained by epistemology. RVC has always been motivated by the claim that the theory of causation should facilitate causal inference. It has also been motivated by the claim that causation has to be the basis for 'recipes and precautions', as Mackie (1974: 141) has put it.

2. REGULARITY AND REALISM

Is RVC a realist theory of causation? If realism implies commitment to (a) the reality and

(b) the mind-independence of the entities one is a realist about, RVC is not inconsistent with realism. Almost all of the defenders of RVC have taken it to be the case that the regularities that are constitutive of causation—the cement of the universe—are real *and* mind-independent: they would exist as (perhaps very complicated) patterns among events even if there were no minds around. One may wonder: in what sense is a regularity real if only some part of it has been, as it were, actualized?

That's a fair worry. A regularity is *not* a summary of what has happened in the past. It is a universality—it extends to the present *and* the future; it covers everything under its fold. So one may naturally ask: what grounds the regularity? What is the truth-maker of the claim that *All As are B*? This is a tough issue and it invites me to swim in waters deeper than I can handle. But it seems that a regularity is best conceived as a perduring entity, since it cannot be said to be wholly present in different regions at different times. That is, a regularity has temporal (and spatial) parts. This view, conjoined with eternalism (the view that past and future objects and times are no less real than the present ones) makes it possible to think of the regularity in a sort of timeless way, *sub specie aeterni*. More specifically, one might think of a regularity as the mereological sum of its instances (that is, its parts), where instances at other times and at other places are temporal and spatial parts of the regularity. Here again, the mereological sum is characterized by the unity of a pattern. The point of all this is that we should distinguish the epistemic question of how we come to know the presence of a regularity, given that our evidence for it has always to do with past and present instances of it, from the metaphysical question of what kind of entity a regularity is.

2.1 Causal Realism

Galen Strawson (1989: 84), defines 'Causation' (with capital 'C') in such a way that to believe in Causation is to believe: '(A) that there is something about the fundamental nature of the world in virtue of which the world is regular in its behaviour; and (B) that that something is what causation is, or rather it is at least an essential part of what causation is' (1989: 84–5). We might think of it as a thick view of causation: there is something—call it *X*—which grounds/explains the regularity; hence causation is: regularity + *X*. If this is what causation is, RVC would fail not because it fails the dual realist commitment noted above (reality and mind-independence), since it does not, but because it leaves out some allegedly essential element of causation. But then, the disagreement between Causation and RVC is not about realism but about what the correct view of causation is.

Is this thick view of causation the right one? Couldn't we get by with a thin view of causation, which dispenses with the extra *X*? Strawson (1987) argues that it is rationally compelling to posit the existence of something other than the regularity to explain the regularity. The basis of this compulsion is what might be called the *ultimate argument against* RVC—what we may also call 'the terminus of explanation' argument: RVC leaves unexplained something that requires explanation, namely, the existence of regularity in nature. Strawson claims that either there is an explanation of regularity or the presence of regularity becomes a matter of chance—a coincidence. It is then alleged that there is need for a deeper

explanation of regularity. This is a popular view, though there is disagreement as to what this deeper explanation should consist in. Some appeal to powers (e.g. Mumford 1998, 2004; Ellis 2001; Molnar 2003), others (including Strawson 1989) posit a force-based productive relation; others appeal to thick laws of nature: that is, laws that are not, ultimately, regularities. (See e.g. Dretske 1977; Tooley 1977; and Armstrong 1983. For a development of the standard criticism of this view, see Psillos 2002: ch. 6. See also Psillos 2006*a*; 2006*b*.)

It's hard to see why a deeper terminus of explanation would be more natural or more preferable. After all, there must be some terminus of explanation—hence, there must be unexplained explainers. Positing an extra layer of ontically distinct facts behind (or below) the regularities will itself be an unexplained explainer. The question 'what explains the regularity' is just pushed back: 'what explains the productive relation (or whatever)?' There is not much gain here, because we should either take the presence of this extra layer of regularity-enforcing facts as self-explanatory or we should just push back the terminus of explanation. As Wittgenstein aptly put it: 'a nothing could serve just as well as a something about which nothing could be said' (1953: §304). And that's exactly what the advocates of RVC should point out: one supposed mystery (the presence of regularity) is not explained by positing another mystery (a supposed productive relation or the like). (See Beebee 2006*b*.) What is more, the positing of an extra *X* like Strawson's productive forces (or some kind of nomic necessitation) does not *eo ipso* yield regularities: there could be powers or forces or whatever without there being any regularity in the world.

The presence of regularities in nature, an advocate of RVC would say, can be explained by appeal to other, more fundamental (and in this sense *deeper*) regularities. So their presence is not a matter of chance. But some regularities, the ultimate and fundamental ones, must be taken as brute: their presence admits of no further explanation. This does not imply that they are a matter of chance. Indeed, admitting that they are a matter of chance would amount to offering a further explanation—a chancy one—of their presence. The friends of RVC firmly deny the alleged need to appeal to a different ontological category (something which is *not* a regularity but has metaphysical bite) to explain the presence of regularities.

It is worth stressing that for an advocate of RVC the key issue is not so much to add something to regularity in order to get causation, but to avoid (and block) the addition of specific ontic features such that they would compromise the fundamental commitment to regularities and their metaphysically irreducible nature. Advocates of RVC would not object (and have not objected) to calls for making the regularity view more robust. But they have persistently argued against the addition of *powers* and other metaphysically heavyweight means to enforce the existence and operation of regularities. An advocate of RVC would view the world as consisting of regularities *all the way down*—this would be its metaphysical blueprint; and yet she would also accept that these regularities are *real and mind-independent*.

Michael J. Costa (1989: 173) has introduced a useful distinction between *causal objectivism* and *power realism*. The former is the view that 'causes are objective in the sense that causal relations will continue to hold among events in the world even if there were no minds to perceive them'. The latter is the view that 'objects stand in causal relations because of the respective causal powers in the objects'. RVC clearly denies power realism. What then of causal objectivism?

Here we need to exercise caution. The regularities that exist in the world are (or can be

conceived of as being) mind-independent in the sense that their existence is independent of the presence of minds: there would be regularities (for example, planets would move in ellipses) even if there were no minds. Yet, *what* causes *what* is not a *fully* objective matter, namely, it is not a matter that is fixed by the world and it alone. Why this is so will become clear in sects. 5 and 6, but the gist is this: on RVC, causation is constitutively dependent on likeness in that it requires that events *like c* (the cause) are followed by events *like e* (the effect). Likeness, though based on objective similarities and differences among events in the world and patterns of dependence among them, is also a matter of respects and degrees of similarity, which are, at least partly, of our own devising. Placing events in similarity classes is the joint product of the world and humans—though it seems that as we go down to the level of fundamental physics, the similarity classes (what we may call *natural* classes) are the product of the world alone.

3. REGULARITY VS. POWER: BROWN VS. REID

The revolt against powers and the concomitant defence of a regularity view found its clearest expression in the writing of the Scottish philosopher Thomas Brown (1778–1820). Brown's main contribution to the philosophy of causation was his book *Inquiry into the Relation of Cause and Effect* (1822).

The intellectual milieu within which Brown operated was dominated by Thomas Reid's power-based account of causation (Reid 1788: Essay 1). Reid spoke freely of active powers and took it that (a) the very concept of power is simple and undefinable; (b) power is *not* something we either perceive via the senses or are aware of in our consciousness (we are conscious only of the *operation* of power and not of the power itself); (c) power is something whose existence we infer by means of reason based on its operation/manifestation; (d) power is distinct from its manifestation/exertion in that there may be unexerted powers; (e) the idea we have of power is relative, namely, as the conception of something that produces or brings about certain effects; (f) power always requires a subject to which it belongs: it is always the power *of* something, the power that something *has*; and (g) causation is the production of change by the exercise of power. Though we are not conscious of powers, Reid insisted that we are conscious of their exertion when our own mental active powers are exercised, as when we decide to raise our hands. Hence, we can conceive of how a cause can exercise its powers because (and only because) we are conscious of how our own active powers are exercised.

Reid was a vocal critic of the view that causation amounts to regular succession. The claim that was to become famous was that Hume's doctrine implies the absurdity that the day is the cause of night and the night is the cause of day because they have constantly followed each other since the dawn of the earth. As Reid characteristically put it: 'Furthermore, when x occurs before y, and x-type events are constantly conjoined with y-type ones, it isn't always the case that x causes y; if it were, Monday night would be the cause of Tuesday morning, which would be the cause of Tuesday night (ibid. Essay 4 ch. 3).

Here is how Brown (1822: p. viii) sums up his own view:

It is most satisfactory, therefore, to know, that the invariableness of antecedence and consequence, which is represented as only a sign of causation, is itself the only essential

circumstance of causation; that in the sequence of events, we are not merely ignorant of any thing intermediate, but have in truth no reason to suppose it as really existing, or if any thing intermediate exist, no reason to consider it but as itself another physical antecedent of the consequent which we knew before.

Brown's motivation for the regularity view was based on the folk epistemic intuition (he would call it a *fact*) that invariable sequence is a sign of causation and in particular on the claim that we would not *call* a sequence of events causal unless it was invariable. This claim, however, is consistent with the further thought that causation has some other essential characteristic in virtue of which it is exemplified in regular sequences of events. Brown's strategy was precisely to demonstrate that regularity is all there is to causation; it 'is itself the only essential circumstance of causation' (ibid.).

This strategy was two-pronged. On the one hand, he developed a series of arguments against powers—advancing what might be called the identity-theory of powers: powers are nothing but the regularity, the uniformity of sequence. On the other hand, he articulated a number of arguments aiming to show 'the sources of various illusions' that have led philosophers to posit powers and to consider causation something more 'mysterious' than regularity. These, according to Brown, include the use of a number of metaphorical phrases used when we think of causation, such as 'connection' and 'bond'. (For more on this, see Brown 1822: Second Part.)

What Brown firmly denied was the idea that between the cause and the effect there is something else (an 'intermediate tie' or an 'invisible bondage') that connects them or binds them together; in particular something of a radically distinct metaphysical nature. Powers, according to Brown, were supposed to be inherent in objects and yet distinct from them; they were supposed to account for the efficiency of causation. According to his identity-theory, 'power is [the] uniform relation [between cause and effect] and nothing more' (ibid. 26). Hence to ascribe a power to an object is to assert that in similar circumstances it will do similar things. This theory is based on a number of arguments, mostly aiming to show that there is no need to posit powers over and above the regularities.

First, powers are mere abstractions (cf. ibid. 19–21). A causal sequence is a concrete sequence of events. It is *causal* in virtue of the fact that it is invariable (it exhibits regularity of order), namely, its antecedent (the cause) has been followed, is followed, and will be followed by its consequent (the effect). When we consider this relation (*this* is always followed by *that*) abstractly, we render the '—is always followed by—' as '—has the power to —'. This move is supposed to unravel the form of causation, namely, what several concrete causal sequences have in common and in virtue of which they are causal. This move, for Brown, is akin to the hypostatization of substantial forms and suffers from exactly the same problem: it converts an abstraction to reality, thereby creating the further problem to explain what this kind of new entity is and does (See also Brown 1851: 35).

Second, powers are the products of double vision (Brown 1822: 28–9). There are substances and they stand in causal relations to each other (that is, in relations of invariable succession). If we knew all these invariable sequences, we would know everything there is to know about what causes what. If we added that these substances have the *power* to produce certain changes, we would not gain any further information about the world. If we thought of power as

distinct from these invariable sequences, it would be possible that we could have information about invariable sequences without knowing a single power.

Third, powers are not needed for the explanation of action (ibid. 5–7). Action amounts to making a difference. An object does not act on anything if its presence or absence makes no difference to anything. But this difference-making can be understood as invariable sequence. Objects that act and are acted upon (that is, causes and effects) are 'truly, in certain circumstances, the reciprocal and immediate antecedents and consequents, in a series of changes' (ibid. 56–7).

Fourth, powers do not explain the regularities. The existence of regularities in nature is not rendered 'less wonderful' by an appeal to powers (cf. Brown 1851: 36).

Brown made an extra effort to neutralize Reid's objection to RVC. He argued that Reid's example of night causing day either does not describe a case of regular and invariable succession or, if it does, it can be fully captured by the regularity view of causation (Brown 1822: 170–1). All depends on how exactly the event-types that are supposed to constitute the regular succession are identified. Given a 'vulgar' (that is, coarse-grained) description of the event-types that are supposed to be in a relation of invariable succession, there is no invariable succession and hence no causation. The night, understood as various degrees of *darkness*, is not invariably followed by day, understood as various degrees of *light*: 'they ... rather appear to follow each other loosely and variously, like those irregular successions of events, which we denominate Accidental' (ibid. 171). Given, on the other hand, a fine-grained description of the event-types, there is regularity and hence causation. Strictly speaking, night and day are not events—they are not even single phenomena, but series of phenomena grouped together by reference to some similarity and difference: degrees of darkness and degrees of light. If we focus on 'the successive *pairs* of that multitude of events, which we denominate night and day' (ibid. 170), and if, further, we take these events to be the positions of the earth in relation to the sun during its rotation around its axis, the motion of the earth immediately before the sunrise does cause the subsequent position of the earth in which the sunlight directly reaches the ground. In this way, the succession of night and day is explained by being reduced to a more complex regularity (picked by a more appropriate description of the causal relata). Brown was fully aware of the fact that an advocate of RVC can claim that an invariable succession between *A* and *B* need not imply that *A* causes *B* or that *B* causes *A*, since *A* and *B* might be the effects of a common cause *C*.

Brown turns on its head the problem raised by Reid. Precisely because regularity constitutes causation, where there is no causation there must be an explanation in terms of the *absence* of regularity; and where there is causation, some regularity must be present, though the grounding regularity need not be described in the vocabulary in which the causal claim is described. Hence, Brown identified the claim that the advocates of RVC should make: the regularities that constitute causation need not be read off directly from the description of events that constitute the relata of a certain invariable sequence; but in so far as there is causation, there is a suitably described underlying regularity. As he nicely put it: 'The generalisations of language are already made for us before we have ourselves begun to generalise.' And this may well lead us 'to suppose a physical relation in many cases where there is none, and to neglect it as often where it truly is' (ibid. note M).

4. CAUSES ARE CONDITIONS: MILL

The programmatic view of RVC, namely, that there is no need or room for a deeper metaphysical story to be told about causation, has been shared by many empiricists. John Stuart Mill put forward the view that there is no difference between cause and antecedent condition in that causes *are* antecedent conditions for effects, and in particular that causes are *sufficient* conditions for their effect. Given that there is, normally, a cluster of factors that constitutes a sufficient condition for an effect, there is no real distinction between the cause and the (standing) conditions among the factors that constitute the invariable antecedent of the effect. Accordingly, causal relations relate several factors *C, F, G,* etc. with an effect *E* such that the conjunction of all these (call it *CFG*) is sufficient for *E*. Following Mill, let's call these factors 'positive conditions'. Strictly speaking, Mill adds, negative conditions, namely, the *absence* of several factors, are also required for the effect *E* invariably to follow. Hence Mill (1911: 217) argues: 'The cause then, philosophically speaking, is the sum total of the conditions positive and negative taken together; the whole of the contingencies of every description, which being realised, the consequent invariably follows.' The real cause is 'the whole of these antecedents' (ibid. 214), namely, the full sufficient condition. It might be objected that the Millian account fails because it allows the inclusion of irrelevant factors in the antecedent condition that was sufficient for an effect *E*. If *CFG* is sufficient for *E*, then so is *ACFG,* where *A* might be totally irrelevant to *E*. But Mill is on safe ground here. There can be many factors coexisting with the causal antecedent of an effect, but they are not part of this causal antecedent because they are not invariably connected with the effect. In effect, the cause should be the minimally sufficient antecedent condition. Another objection might be that Mill's denial of the difference between causes and conditions might lead him to accept *trivially* relevant causal factors. Suppose that a person died *after* drinking arsenic. Why shouldn't we include in the conditions of her death the fact that she was human and not, say, a robot, or the fact that she was a woman and not a man, or indeed the fact that she was alive before her death? Here too, a Millian can accept Mackie's (1974: 63) notion, implicit already in Mill (1911: 214–15) of a 'causal field'. This is the *context* in which the conditions of an effect occur. The causal field should be taken to be the *background* 'against which the causing goes on' (Mackie 1974: 63). This background would be there even if the specific conditions that are sufficient for the occurrence of the effect were absent.

Like Brown before him, Mill tried to meet Reid's counterexample head-on. But unlike Brown, he thought that some new condition is called for if RVC is to meet this counterexample. According to Mill, regular association (or invariable succession) is not sufficient for causation. What must be added to invariable succession to get causation is 'unconditionality'. Mill does not explain this notion in great detail, but what he has in mind is that for *B* to be the effect of *A* it is not enough for *B* to follow *A* invariably, as a matter of fact; it is also necessary that *B* follows *A* under *any* circumstances. Hence, the dependence of *B* on *A* should be such that it is not conditional on the presence of other factors—say *C*—which are such that they are sufficient for *B*: given *C, B* follows irrespective of whether or not *A* is present. Mill's (1911: 222) reply to Reid was that day does not follow night unconditionally. As he characteristically put it:

There are sequences, as uniform in past experience as any others whatever, which yet we do not regard as cases of causation, but as conjunctions in some sense accidental. Such, to an accurate thinker, is that of day and night. The one might have existed for any length of time, and the other not have followed the sooner for its existence; it follows only if certain other antecedents exist; and where those antecedents existed, it would follow in any case.

A clear case in which there is no unconditionality is when two events are joint effects of a common cause (ibid. 252–3). Though there is invariable succession, event E_1 (or E_2, for that matter) would follow given the presence of the cause C irrespective of the occurrence of the other effect. A clear case in which unconditionality is ensured is when the cause is also necessary for the effect. Unconditionality is, for Mill, the valid residue of the traditional claim that there is necessity in nature. According, then, to his version of RVC (ibid. 222), 'We may define, therefore, the cause of the phenomenon to be the antecedent, or the concurrence of antecedents, on which it is invariably and *unconditionally* consequent.'

Mill takes it that a sequence of events is unconditional if it falls under a law of nature. Laws of nature capture the valid residue of the traditional conception of necessity: 'That which is necessary, that which *must* be, means that which will be, whatever supposition we may make in regard to all other things' (ibid. 222).

It can then be said that for Mill, the correct statement of RVC is something like this:

> M-RVC
>
> c causes e iff

     i. c is spatiotemporally contiguous to e;

    ii. e succeeds c in time; and

   iii. *it is law of nature* that all events of type C (i.e. events that are like c) are regularly followed by events of type E (i.e. events like e).

5. REGULARITIES AND LAWS OF NATURE

The view that laws of nature are regularities can be called the *Regularity View of Laws* (RVL). This is meant to be a metaphysical thesis about lawhood: the worldly stuff that laws consist of is regularities. RVL denies that laws, as they are in the world, are anything over and above stable patterns of events. Programmatically, RVC ties causation to the presence of regularities: to call a sequence of events c and e causal is to say that this sequence is a part of (instantiates) a regularity, namely an invariable (and unconditional, according to Mill) succession between event-types C and E. But not all regularities are fit to capture causal connections. Let us follow Mill (and customary usage) and call *accidental* (or accidents) those regularities that are not laws of nature. Then, the proper statement of RVC should be taken to be:

causation is lawlike regularity.

RVL asserts that:

laws of nature are regularities.

If laws are *merely* regularities, there is no distinction between law and regularity; then, *all* regularities are lawlike; hence causation is *merely* regularity. But this is exactly what Mill's version of RVC (M-RVC) has aimed to avoid. Hence, laws of nature should be regularities plus something else, something sufficient to distinguish an accidental regularity from a lawlike one. RVL* is the thesis that

laws of nature are regularities + *Y*.

RVL* shares with RVL the basic metaphysical commitment to regularity. But one should be careful here. Staying within the bounds of a Humean view of causation depends on what the differentia *Y* is. More specifically, it should not be anything such that it introduces (or implies) any metaphysically distinct, and deeper, kind of entity, anything like powers or potencies and the like that are meant to ground or explain the regularity. Nor should it imply commitment to any kind of natural necessity (or relation of necessitation) that is repugnant to Humeans. Let's call this extra *Y* the property of lawlikeness. What can it be?

There have been a number of candidates. (For more detailed discussion see Psillos 2002: ch. 5.) But the most promising attempt is the *web of laws* view. According to this view, the regularities that constitute the laws of nature are those that are expressed by the axioms and theorems of an ideal deductive system of our knowledge of the world, and in particular, of a deductive system that strikes the *best* balance between simplicity and strength. (If there is no unique best deductive system, the laws are expressed by the axioms or theorems that are common to all deductive systems that tie in terms of simplicity and strength.) Simplicity is required because it disallows extraneous elements from the system of laws. Strength is required because the deductive system should be as informative as possible about the laws that hold in the world. Whatever regularity is not part of this *best system* it is accidental: it fails to be a genuine law of nature. The gist of this approach, which was advocated by Mill, and in the twentieth century by Ramsey (1928) and Lewis (1973*a*), is that no regularity, taken in isolation, can be deemed a law of nature. The regularities that constitute laws of nature are determined in a kind of holistic fashion by being parts of a structure.

The Mill–Ramsey–Lewis view has many attractions. It solves the problem of how to distinguish between laws and accidents. It shows, in a non-circular way, how laws can support counterfactuals since it identifies laws *independently* of their ability to support counterfactuals. It makes clear the difference between regarding a statement as lawlike and its being lawlike. It respects the major empiricist thesis that laws of nature are contingent: a

regularity might be a law in the actual world without being a law in other possible worlds, since in these possible worlds it might not be part of the best system for these worlds. It solves the problem of uninstantiated laws: these are proper laws in so far as their addition to the best system results in the enhancement of the strength of the best system, without detracting from its simplicity.

The best candidate, then, for RVL* is the Mill–Ramsey–Lewis view. Laws are regularities (this is their worldly stuff) and the differentia Y is 'being expressed by an axiom or theorem in the best deductive system'. It seems, however, that this differentia cannot be fully objective. That a statement is implied or not within a deductive system is an objective matter, something that obtains independently of our knowledge of it. But what statements are (objectively) implied depends on the *way* the deductive system is organized, something that is not necessarily objective (in that there may be a lot of freedom in picking the axioms). Even if we fixed the slippery notion of simplicity, there seems to be no objective way to strike a balance between simplicity and strength. Nor is it guaranteed that there is such a balance. Hence, what regularities will end up being *laws* is based, at least partly, on epistemic criteria and, generally, on our subjective desideratum to organize our knowledge of the world in a deductive system.

There is something to this objection, but it should not be overstated. The worldly stuff that laws consist of are regularities and they are not mind-dependent: they characterize the world irrespective of our knowledge of them and of our being able to identify them. The feature, however, that renders some regularities laws at the expense of some others (the accidents)— the property of lawlikeness—is *not* worldly. It is broadly subjective, though not arbitrary. It seems that this is a price the Humean has to pay in order to avoid certain metaphysical commitments.

The repercussion for causation is obvious. RVC (in its sophisticated Millian version M-RVC) takes it that causation is *lawlike* regularity; if the lawlikeness of a regularity is not something fixed by the world (even if the regularity is), RVC cannot be fully objective—it rests on a broadly subjective (but not arbitrary and whimsical) circumscription of the regularities that constitute causation. The regularities themselves are mind independent, but what causes what is not.

There is another route to the same point, which is relevant to what follows. As Nelson Goodman (1983) has shown using the famous predicate 'is grue' (defined as: observed before 2010 and green or not observed before 2010 and blue), the very idea of lawlikeness requires a theory of what *predicates* can be constituents of lawlike statements. Quine (1969) and others (including Goodman) took it that the predicates suitable for lawlike statements (statements that express laws of nature) must pick out natural kinds. Exactly the same need arises in connection with the Mill–Ramsey–Lewis view of laws. It is perfectly possible that the simplest and strongest deductive systematization may be effected by 'unnatural' predicates, that is, predicates that do not pick out natural kinds. Then all sorts of odd regularities would end up being laws, since they would be captured by axioms or theorems of the 'best system'. All this means that the prospects of a theory of lawlikeness are tied with the prospects of a theory of natural-kind predicates.

How exactly a natural kind is circumscribed is not a very straightforward issue (at least for someone who is not an essentialist) but the least that is involved in the characterization of a

kind of entities is that they are *like* each other in relevant degrees and respects. What respects of likeness are relevant to kind-membership? Here, the obvious answer would be: those respects in virtue of which entities have similar nomological and causal behaviour. This would create an air of vicious circularity since it seems that for a regularity to be a law it should constitute a pattern among natural kinds (expressed by 'natural kind' predicates) and conversely for a kind to be natural it should part of a nomological pattern. But this kind of circularity seems inevitable. There is an intimate connection between the issue of what laws of nature are and the issue of what kinds are natural: one cannot be delineated without the other.

Some conception of similarity becomes necessary in thinking about which regularities are laws. As we have already seen, this very idea of classes of resembling events enters constitutively into RVC. If similarity is not a fully objective relation—in that the respects and degrees of similarity are not fixed by the world—this is another entry-point for subjectivity into causation. Perhaps, as we go down to the level of atoms and elementary particles and their properties, this element of subjectivity is diminished. At that level, it seems there are fully objectively circumscribed natural classes: not only are the similarities and differences between types of elementary particles objective, but also the respects and degrees in which they are similar and different are fixed by the way the world is.

Actually, one might want to follow Lewis (1984; 1986*a*: 50–2) and posit the existence of natural properties (or classes) that are distinguished from each other by objective sameness and difference in nature. Natural properties, it might be said, carve nature at its joints; they provide an objective classification—hence they are fully objective. There is a lot to be said for this view, and it shows the way an advocate of RVC has to go if she wants to avoid buying into a lot more subjectivism about causation and lawhood than she is willing to accept. This notion of naturalness is hard to define—Lewis takes it as an unanalysed, yet indispensable, primitive. Indeed, and for our purposes, if we try to explain similarity as sharing of natural properties, then we cannot analyse naturalness in terms of (objective) similarities. According to Lewis, the inventory of perfectly natural classes (properties) is (or will be) delivered by fundamental physics. But even allowing for perfectly natural properties, naturalness, as Lewis himself argues, is a matter of degree. Some properties are more natural than others (say, mass or charge relative to colours and colours relative to gruesome properties). But the very idea of degrees of naturalness implies that as we move away from the 'perfectly natural' end of the natural–unnatural continuum and towards the 'highly unnatural' (disjunctive, gerrymandered) end of it, similarity judgements become less and less objective and more and more dependent on us (our categories and classificatory schemes).

6. REGULARITY AND SIMILARITY: VENN

Causation implies similarity. This is something already present in Hume's first definition of causation, we may recall. What John Venn saw clearly was that this dependence on similarity has important repercussions for RVC. In its sophisticated (Millian) version, RVC takes the antecedent of a sequence to be the complete cluster of factors that constitute a sufficient and a necessary condition for the effect. Venn added that the full statement of the regularity should in fact be a disjunction of conjunctions such that it is necessary and sufficient for the effect. Given that the cause is a conjunction (or cluster) of factors and that

there can be a plurality of causes for a certain effect, the regularities should be captured by the following logical form:

(*ABC*) or (*DFG*) or ... ↔ *E*.

This way of putting the regularity implies that the effect is an event-type, devoid of its individuality. If the effect were fully specified in all its detail, its cause would have been a complex concrete event too—but then hardly any repetition would be possible. It would be *this* causing (or being followed by) *that*. The move from *this* causes (or is followed by) *that* to *this-type* causes (or is followed by) *that-type*, a move which for Venn is necessary for the very possibility of inductive inference, relies on classifying events under similarity classes. As Venn put it, 'No two objects or events in nature are alike in all their details, and therefore if we want to secure repetition we must submit to let go some of the characteristics' (1889: 57). Differently put, that there are laws that cover causal sequences of events follows from the fact that there are similarities among the objects/entities/events that are involved in these sequences. If there were no similarities, if each sequence were unique, the very fabric of inductive inference would be disrupted.

This reliance on similarity (*events of type C are followed by events of type E*) introduces, according to Venn, a subjective element to causation. Similarity has an objective and a subjective aspect. The crafting of the event-types whose joint recurrence marks a regularity is done jointly by nature and us in that it is up to us, in the final analysis, which of the elements that compose the antecedent of a sequence must be omitted, abstracted away and the like so that the repetition of the thus-constituted antecedent event-type is safeguarded. As Venn (1889: 98) nicely put it:

Such repetitions as we actually find set before us are the results of two factors, one contributed by nature the other partly contributed by ourselves. ... Nature ... as Leibniz was fond of insisting, never exactly repeats herself. But she does the next best thing to this for us. She gives us repetitions—sometimes very frequent, sometimes very scarce, according to the nature of the phenomena—of all the important elements, only leaving it to us to decide what these important elements are.

A more systematic way to put Venn's problem is this. If the reference-classes of the causal relata were unit classes, we would have absolutely precise and (trivially) exceptionless causal claims, but scarcely any repetition. If, on the other hand, the reference-classes in which the causal relata belong were broad, there would be repetition but the relevant causal claims would be less precise and not necessarily exceptionless. RVC does not tell us how the reference classes (that is, the similarity classes) in which the causal relata are put are picked; for RVC to work there must be a way (or a theory) to do this. As Arthur Pap (1952: 660) pointed out, RVC faces a dilemma: either all sequences are coincidental (since, if the causal relata are specified in a coarse-grained way, the sequences are not exceptionless) or all sequences are causal (since, if the causal relata are unit-classes of events, all sequences are trivially exceptionless).

To avoid this dilemma, RVC must find the golden mean between a very fine-grained and a

very coarse-grained description of the causal relata. The events whose pattern of joint recurrence constitutes a regularity should belong to similarity classes that are neither too broad to allow exceptions nor too narrow to bar repetition. As noted at the end of the last section, there is need to appeal to natural classes. Perhaps, the needed golden mean is achieved primarily at the level of elementary particles and their interactions. At this level, it can also be plausibly said that, by virtue of being perfectly natural, the similarity classes are fully objective.

7. REGULARITY VS. SINGULARITY: DAVIDSON

What then is the correct statement of the Regularity View of Causation? The answer to this question has found its *locus classicus* in Donald Davidson's writings, though the gist can be found in Pap (1952). The key thought, which motivates Davidson's view too, is that RVC does *not* offer a recipe for (or a rule of) translation of singular causal statements into general causal ones. Take any singular causal claim, for example '*c* caused *e*'. The aim of RVC is not to translate this singular statement into a general one. Rather, RVC is committed to there being a law such that events described as events of type *C* (where *c* is one of them) are followed by events described as events of type *E* (where *e* is one of them). The existence of the law is assured, but its description (and its exact statement) does not directly follow from the descriptions used to identify the relata of the singular causal statement.

The gist of RVC, we have seen, is that a sequence of events is causal only by right of its membership in a class of similar sequences. This is important because though the causal relation seems to have the same surface structure with other relations, its deep structure is vastly different, if RVC is right. On RVC, whether or not a particular sequence of events (*this* billiard ball moving *that* billiard ball after colliding with it) is causal depends on things that happen elsewhere and elsewhen in the universe, and in particular on whether or not this particular sequence instantiates a regularity. It depends, that is, on whether event-tokens *c* and *e* fall under suitable event-types *C* and *E* such that all events of type *C* are regularly associated with (or, regularly followed by) events of type *E*. '*c* causes *e*' has the same structure as '*x* loves *y*' (or '*x* is taller than *y*'); but, it would be absurd to say that whether or not 'Mary loves John' is true depends on anything other than Mary, John, and their (local) properties and relations. This, of course, is another way to put the claim that causation is extrinsic to its *relata*: it depends on *general* facts; on what happens at other places and at other times.

A different view has emerged in the twentieth century, mostly in the writings of Curt John Ducasse. This, like RVC, is metaphysically lean. It does not posit powers and the like to explain causation. It differs from RVC in taking causation to be a singular relation, fully captured by whatever happens there and then between two concrete events in their full individuality. On this singularist view, if there were no repetition in nature, there would still be causation, in so far as there were change in nature. Causal laws may well exist in nature, but only because there are causal facts in their own individual right; (causal) laws are generalizations over causal facts and not (as RVC would have it) constitutive of causal facts.

We shall not go here into the details of this view. But some very general points are important for the proper defence of RVC. According to Ducasse's (1951, 1968) account, an

event *c* caused an event *e* if and only if *c* was the only difference in *e*'s environment before *e* occurred. To put it more precisely, suppose we have a concrete state of affairs *S* (*a, b, c, d, e, f*) and a (single) change *C* (*a, b*) of features *a* and *b* of *S* which is followed by a change *C* (*e, f*) of features *e* and *f* of *S*. Then, *C* (*a, b*) was the cause of *C* (*e, f*). As Ducasse (1951: 108) put it, 'a cause is always a *difference* occurring in a state of affairs *S* in which the effect is another and later *difference*'. There are obvious problems with the epistemology of this view, but leaving them to the one side, we may ask: how is this very idea of *difference* to be understood? Similarly, what changes (that is, differences) are relevant to the effect? To answer these questions, there is need to ascend to the level of event-types, that is, to descriptions of the events not in their full concrete individuality (whatever that means!) but to descriptions that bring out in what respects events have changed or remained the same.

When the baseball shattered the window, the canary in the nearby cage had just started singing. In its full concrete individuality, the cause of *this* breaking of *this* window was the full antecedent state (including the airwaves of *this* canary song hitting the window at the moment *this* baseball hit it). What Ducasse does not seem to appreciate is that the very idea of difference, even when it comes to difference in events in their concrete individuality, requires appeal to general facts of similarity and difference. He notes that when it comes to concrete events in their full determinateness, we can only use designators such as 'whatever is occurring here and now' and the like (1951: 152–3). But, of course, for whatever-is-occurring-here-and-now to be considered as a difference (or as a change) it has to be classified in certain way, leaving out some part of it (whatever-remains-the-same) and focusing on whatever has changed (from being some way to being some other way). To consider a concrete event *qua* an instance of an event type is to concede that there are general patterns under which events fall; and if this is the only way to make sense of differences and similarities among events, the very idea of causation as difference-making (*changes*) implicates general—and not singular—facts.

A version of the point above has been stressed by Davidson (2005). Davidson aims to show how *both* Humeans and singularists need to rely on similarity in their accounts of causation; Humeans in order to say when events are similar to each other and singularists (Ducasse in particular) in order to say when events are relevantly *dissimilar* to each other. According to Davidson, in order for Ducasse to claim that a *c*-change caused an *e*-change, there must be event types *C'* and *C* (and event-types *E'* and *E*) such that the change of *c* from being *C'* to being *C* caused the change of *e* from being *E'* to *E*, that is, there must be a *c*-like event (meaning an event-type of the form: *has changed from C' to C*) and an *e*-like event (meaning an event-type of the form: *has changed from E' to E*) such that *c*-like events are followed by *e*-like events.

Davidson (2005: 212) takes all this to show 'that singular causal statements imply the existence of covering laws: events are changes that explain and require such explanation'. This is simply to reinforce his old and well-known point that all causation is nomological.

As is well-known, Davidson (1967) argued that there is room for reconciliation between the singularist and the Humean. When we pick the descriptions of the events that enter a causal statement, the descriptions may be such that they entitle us to *deduce* the singular causal statement from a lawlike statement together with the assumption that the events referred to in the statement occurred. So we can subsume the singular causal statement under a causal law. His suggestion (ibid. 83) is that if '*c* causes *e*' is true, there must be descriptions of events *c*

and *e* such that they fall under a law from which it follows that the first event caused the second, even if this law is unknown to those who use the singular causal statement, and even if the law is not stated in the vocabulary of the singular causal statement.

Davidson has managed to bring together (and in line) the two key points on behalf of RVC made by Brown and Venn. Brown noted that though where there is causation, there is regularity, the underpinning regularity might be much more complex than the one implied by a face-value reading of the causal claim: there are descriptions of the events that constitute the causal sequence such that they fall under a regularity. Venn noted that what goes into the description of events so that repetition is ensured relies on judgements of similarity. Davidson's point is that precisely because of this there will be laws that underpin causal assertions—causation is always a matter of law.

8. REGULARITY AND EXPLANATION

The idea that causes are nomologically sufficient for the effects was the kernel of the Deductive-Nomological model of explanation (henceforth, DN-model), advanced by Carl Hempel and Paul Oppenheim (see Hempel 1965; for details of the DN-model and a qualified defence of it, see Psillos 2002: ch. 8). According to it, a singular event *e* (the *explanandum*) is explained if and only if a description of *e* is the conclusion of a valid deductive argument, whose premises, the *explanans*, involve essentially a lawlike statement *L*, and a set *C* of initial or antecedent conditions. The occurrence of the *explanandum* is thereby subsumed under a natural law. Schematically, to offer an explanation of an event *e* is to construct a valid deductive argument of the form:

    (DN)
  Antecedent/Initial Conditions *C*1, ..., *Ci*
  Lawlike Statements *L*1, ..., *Lj*

  ―――――――――――――――――――――

  event/fact to be explained (*explanandum*) *e*

Hempel took his model to provide the correct account of causal explanation. As he put it, 'causal explanation is a special type of deductive nomological explanation' (1965: 300). This does not imply that all DN-explanations are causal. The thesis is that all *causal* explanation is DN-explanation, which means that for *c* causally to explain *e* it should be the case that there are relevant laws L_1, ..., L_n in virtue of which the occurrence of the antecedent condition *c* is nomologically sufficient for the occurrence of the event *e* (cf. ibid. 349). In elaborating this view, Hempel (ibid. 350) noted that when we say that event *c* caused event *e*, 'the given causal statement must be taken to claim by implication that an appropriate law or set of laws holds by virtue of which [*c*] causes [*e*]'. Thus put, the claim is not far from Davidson's (1967: 84) point made in the previous section: true statements of the form '*a* caused *b*' imply commitment that 'there are descriptions of *a* and *b* such that the result of substituting them for "*a*" and "*b*" is entailed by true premises of the form [the relevant law] and [initial

conditions]'.

According to both Hempel and Davidson, causation is a matter of nomological sufficiency: *c* causes *e* iff there is a *law* that connects events like *c* with events like *e*. But, being interested in causation, Davidson denied that the covering law should be specified or searched for. Being interested in explanation, Hempel insisted that the causal explanation is incomplete unless the DN-argument is fully spelt out and the nomological statement is made explicit. There cannot be a deduction of the *explanandum* from the *explanans* (and hence an explanation of it) unless the *explanans* is fully specified and is such that at least one law is stated. This simply means that the *explanandum* (the effect) and the *explanans* (the cause and the law in virtue of which it operates) must be described in shared vocabulary: the law should ultimately be read off from the description of the causal relata (perhaps with some help from bridge principles). Hempel thought, rightly, that when the law is not explicitly offered in a causal *explanation,* the explanation is incomplete. It is like, as he said, being given 'a note saying that there is a treasure hidden somewhere' (1965: 349). But then again, the treasure might turn out to be fool's gold if the generalization that is supposed to do the explaining is not a proper law. The problem that Davidson identified is precisely that there is no simple and straightforward way to go from a singular causal statement of the form '*c* caused *e*' to a general statement that is a covering *law.* The DN-model oversimplifies this move. Even accidentally true generalizations can play the 'covering' role. So there is a need for a sharp separation between those regularities that are laws and those that are accidents. Besides, there can be non-strict, non-exceptionless generalizations that are explanatory nonetheless, even though they cannot function as premisses in a DN-argument. *Ceteris paribus* generalizations can, that is, be explanatory.

Laws constitute the link between RVC and the DN-model of explanation. On RVC, where there is causation, *there are* covering laws; and yet, what exactly these laws are is a different matter. On the DN-model of explanation, this different matter does matter. Causal explanations are arguments and for the argument to be deductively valid and explanatory it is required that (descriptions of) laws should be among its premisses. The *explanandum* and the *explanans* must share vocabulary. However, the laws that render a sequence of events causal need not be captured by the lawlike statements in virtue of which an explanation of this sequence of events proceeds. For instance, the law that all metals expand when heated may be plausibly used to explain why a certain piece of iron expanded when it was heated, and yet it may be the case that the causal law that underpins this sequence of events is too complex to be neatly captured by the statement 'All metals expand when heated.'

9. COMPLEX REGULARITIES: MACKIE

Singular causal statements do not imply specific lawlike statements—if the sophisticated Davidsonian version of RVC is right, what follows is that *there is* a law under which they fall. But singular causal statements do not imply the presence of a productive relation either. From the ordinary meaning of singular causal statements we cannot draw any conclusions about what causation is. They are simply neutral on this matter.

John Mackie (1974) suggested that regularities do play some role in causation as it is in the world in the sense that singular causal statements are *grounded* in regularities (even though they do not imply any regularities). Mackie actually held no brief for RVC. He thought there is a lot more to causation than regularity. But he did emphatically deny that this more would imply anything like substantive commitments to powers or necessary connections or nomic universals. The attraction of the regularity view is precisely that 'it involves no mysteries' (ibid. 60) and that it makes vivid how causal facts can be known and how causal inference works.

What kinds of regularity ground causation? Mackie argued that these are *complex* regularities. Typically, effects have a plurality of causes: a certain effect can be brought about by a number of distinct clusters of factors. Each cluster is sufficient to bring about the effect, but none of them is necessary. A house, for instance, catches fire and gets burned to the ground. There are a number of clusters of factors that can cause house-fires. One cluster includes the occurrence of a short circuit along with the presence of oxygen, the presence of inflammable material in the house, the absence of a sprinkler system, and so on. Another cluster includes the presence of an arsonist, the use of petrol, the presence of oxygen, and so on. Yet another includes the eruption of fire in a neighbouring house, etc. Each cluster is a logical conjunction of single factors. The disjunction of all such clusters (conjunctions) captures the plurality of causes. Each conjunction of factors is sufficient for the fire, but none of them is necessary, since another conjunction of factors can be sufficient for the fire. To simplify matters a little, let us suppose that the regularity has the form:

AX or $Y \leftrightarrow E$,

where AX and Y are clusters of factors that are minimally sufficient for E. To say that AX is minimally sufficient for E is to say that AX is sufficient for E and that none of its conjuncts (A and X) are redundant: none of them, taken on its own, is sufficient for E. The conjunction AX, however, is not necessary for E. E can occur if Y occurs. Each single factor of AX (A, for example) is related to E in an important way. It is, as Mackie has put it, an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition for E. Using the first letters of the italicized words, Mackie has called such a factor an *inus* condition. Causes, then, are at least *inus* conditions.

Causes are *at least inus* conditions. On Mackie's version of RVC, causes can also be either sufficient conditions or necessary conditions, or both. A causal regularity can have any of the forms:

     i. $A \leftrightarrow E$

    ii. $AX \leftrightarrow E$

   iii. A or $Y \leftrightarrow E$

   iv. AX or $Y \leftrightarrow E$.

Of these forms, only (iv) has A to be an *inus* condition for E. According to (i), A is a sufficient and necessary condition for E; according to (ii) A is an insufficient but necessary

part of a sufficient and necessary condition for E; and according to (iii) A is a sufficient but not necessary condition for E. Mackie's improved version of RVC entails that the generic claim to which this view is committed is this:

> For some X and for some Y (which may, however, be null), all (AX or Y) are E, and all E are (AX or Y) (cf. ibid. 71).

This view has a lot of merit. It shows how RVC can deal with exceptions. If there was a short-circuit but there was no fire, it must be because some of the other conditions that were necessary for the *inus* condition to cause the effect were not present. It also shows how causal inference can work. If we know that an effect of the type E has occurred, and if we also know that the set of factors Y was not present, we can conclude that the set of factors AX was present and, in particular, that A was present. More importantly, Mackie's approach allows '*elliptical* or *gappy* universal propositions' (ibid. 66). Suppose the regularities in the world are of the complex disjunction-of-conjunctions type that Mackie (and Venn) have envisaged. Our knowledge of these regularities will be, for the most part, gappy or incomplete. If Mackie's version of RVC is correct, there must be a full universal proposition that completes the gappy or elliptical one. The more the latter is filled in, the more we know about the full complex regularity.

10. ASYMMETRIES

Among the problems faced by RVC, even in Mackie's sophisticated version, two stand out. The *first* is that it fails to distinguish between genuine causes and mere joint effects of a common cause. Mackie (ibid. 84) discusses in some detail Russell's well-known example in which the sounding of the hooter in Manchester is an *inus* condition for workers in London leaving their work. The structure of a counterexample of this sort is this: suppose there are two effects E_1 and E_2 such that they both have C as an *inus* condition: CX or Y is necessary and sufficient for E_1 and CZ or W is necessary and sufficient for E_2. The complex condition E_1&not-Y&Z is sufficient for E_2 and the complex condition ((E_1&not-Y&Z) or W) is necessary and sufficient for E_2. So, E_1 is an *inus* condition for E_2. More generally, it may well be that an effect of a cause can be part of a set of sufficient conditions for another effect of the same cause. Mackie saw in this type of problem a reason that the regularity view is far from being a complete account of causation as it is in the world. His remedy was based on the notion of causal priority (which is not the same as temporal priority).

One interesting suggestion, on behalf of RVC, has been that spurious correlations can be seen as implicating some nomic connection between the correlated events which, however, is *less direct* than the nomic connection between a common cause and the correlated events. This thought has been explored by John F. Clendinnen (1992). How exactly to understand the idea of a more direct nomic connection is complicated. An example might help. The drop in the barometer is correlated with the subsequent storm, but does not cause it—the common cause of both is a drop in the atmospheric pressure. In this set-up, it is clear that the barometer–storm (B–S) nomic connection is *less* direct than the pressure–storm (P–S) nomic connection.

The P–S connection is there even if the B–S connection is not (as an experiment can testify). The P–S connection can be subsumed under more basic laws without this being dependent in any way on the B–S connection. The scope of the P–S connection is greater than the scope of the B–S connection. This leaves us, among other things, with the need to explain when a law is more basic than another. We are not totally in the dark here—the M-R-L view (see sect. 5) can offer some help.

There is no straightforward way out of the joint-effects problem for RVC. If, for instance, we appealed to the absence of direct causal processes between the joint effects, or the absence of counterfactual dependence among them, the solution would be much neater. But the situation is not entirely hopeless for RVC. There will be cases in which there is indeterminacy as to what causes what and, occasionally, there will be joint-effect structures that may pass as causal. But many of them will be resolvable into more complex patterns of nomic dependence.

The *second* problem is that the very idea of causation as having to do with necessary and sufficient conditions blurs the distinction—that is, the asymmetry—between cause and effect. This has led many (including Lewis 1973*b*) to claim that RVC is almost beyond repair. The problem of the direction of causation is vexing and intricate and has had no fully satisfactory solution so far on the basis of *any* account of causation. Hence the fact that RVC cannot solve it should not be taken as a fatal blow against it. The standard answer is that the causal direction is simply the temporal direction: causes *precede* effects in time. This is taken to be problematic because (a) it blocks backward causation on a priori grounds and (b) it lands in a circle, if the direction of time is analysed in causal terms. Neither objection seems to be fatal. Be that as it may, could there be a more informative answer on behalf of the advocate of RVC?

Let us think for a minute of the way Lewis (1973*b*) tries to solve this problem within his counterfactual theory of causation. There is an asymmetry between the past and the future: the former is *fixed*, whereas the latter is *open*. This asymmetry is accounted for in terms of the asymmetry of counterfactual dependence. The past is 'counterfactually independent of the present', since it would remain the same whatever we did now. But the future is not. It depends counterfactually on the present: on what we do now. Lewis argues that this asymmetry of counterfactual dependence is the result of a *contingent* fact, namely, that every event is excessively overdetermined by subsequent events, but very few events are overdetermined by their history. Couldn't a parallel point be made on behalf of RVC? To simplify matters, suppose that the complex regularity has the form:

AX or $Y \leftrightarrow E$

This exhibits an interesting asymmetry. The effect does not determine which disjunct was present, but each disjunct is sufficient for the effect. Because of the complex form of regularities, the causes 'overdetermine' their effect but not vice versa. Here again we talk about a contingent fact, namely, that regularities have this complex form. The cause and the effect are nomologically dependent on each other, but there is an 'inferential' asymmetry: the presence of the effect can be inferred provided some of its causes are there, while the presence of a specific cause cannot be inferred from the presence of the effect. (A version of this reply is in Baumgartner 2008.) This move will not work if regularities do not have this complex form—in such cases, there can be no judgement of asymmetry. It may be objected this 'inferential' asymmetry is not metaphysically robust in that there is no reason why the world

should consist of complex regularities. The friend of this move can reply that though this may well be so, the inferential asymmetry is still important precisely because it puts a non-trivial constraint on causal claims—where it fails, there can be no causal attribution and/or explanation. (Interestingly, Hausman (1998) has aimed to account for the asymmetries of causation in terms of a condition of independence that holds contingently, namely that every event that has any cause has at least *two* independent causes. As Hausman (ibid. 64) notes this kind of condition should at least be taken to impose a necessary condition for the possibility of making causal claims and for offering causal explanations. If it does not hold, no claims of causal asymmetry can be made.)

11. CONCLUDING REMARKS

To the best of my knowledge, there is no theory of causation that is free of counterexamples. Nor is there any theory of causation that tallies best with all our intuitions about what causes what. Nor are these causation-related intuitions always clear-cut and forceful. Perhaps, this is reason enough to make us sceptical about the prospects for a single and unified metaphysical account of causation—of what causation *is* in the world. Perhaps, what we are trying to figure out—*causation*—is not one single condition with a determinate nature. I have tried to substantiate this sceptical stance and to advance causal pluralism (see Psillos 2008). Be that as it may, we can only engage in a cost–benefit analysis of the several competing theories (and perhaps rely on our own intuitions and epistemological preferences). To me then, RVC is the next best thing. A full appraisal of the prospects of RVC would require a more systematic comparison with all other metaphysically thick theories of causation than can be given in this chapter. The important thing, it seems to me, is that RVC is metaphysically lightweight while going a long way towards explaining the epistemology of causation.

As I noted already, it would be wrong to think of RVC as being a rival to other accounts that take causation to be a robust relation of dependence between discrete events. The substantive rivalry is between dependence accounts of causation and production ones. Why, then, couldn't a sensible defence of RVC appeal, say, to counterfactual conditionals? Would the employment of counterfactuals, say in dealing with the problem of joint effects, imply that RVC has been abandoned? In a very strict philosophical sense, the answer is positive—we would appeal to conceptual resources that seem to fall outside the scope of RVC. In a more liberal sense, the appeal to counterfactuals could be seen as a legitimate move, since it does not put in danger the metaphysical agenda of RVC (and its key denial of an extra layer of regularity-enforcing facts). In a similar fashion, the existence of less-than-perfect regularities (say, statistical regularities) is not detrimental to RVC. Probabilistic causation can actually be seen as a completion of RVC—an extension of the key ideas to stochastic phenomena. Here again, the important thing is that the programmatic metaphysics of RVC is not compromised.

In all probability, RVC will need a little help from its friends (that is, other dependence accounts of causation) to sustain its rivalry with its enemies (production approaches). Still, the message is clear: causation does not need a thick metaphysical underpinning.

The Regularity View of Causation is not currently very popular among philosophers, so it is hard to find recent papers and/or books that have mounted serious defences of it. Criticism, on the other hand, abounds. Fair critical presentations of the basic tenets of RVC can be found in Paul Horwich, *Asymmetries in Time* (1987: ch. 8) and Daniel Hausman, *Causal Asymmetries* (1998: ch. 3). Perhaps the classic statement and critique (almost a qualified defence) is J. L. Mackie, *The Cement of the Universe* (1974), especially ch. 3. David Lewis, 'Causation' (1973*b*) presents some central problems that RVC faces as a way to motivate his own counterfactual approach. But Helen Beebee, 'Does Anything Hold the Universe Together?' (2006*b*), has persuasively argued that Lewis's counterfactual theory can be seen as a sophisticated version of the regularity view. Beebee's paper also presents a sophisticated and creative defence of some key metaphysical features of the regularity view. Michael Baumgartner, 'Regularity Theories Reassessed' (2008), makes some headway in the technical articulation of RVC. Clendinnen, 'Nomic Dependence and Causation' (1992), is a suggestive defence of nomic dependence accounts of causation. Chapters 1 and 2 of my *Causation and Explanation* (2002) present my own account of Hume's version of the regularity view and its development by Mill, Mackie, and others. James Woodward, *Making Things Happen* (2003), offers a systematic analysis and criticism of many competing theories of causation and explanation. The most philosophically sophisticated attack on the metaphysics of RVC can be found in Galen Strawson, *The Secret Connexion* (1989). Several entries on theories of causation that appear in the *Stanford Encyclopedia of Philosophy* (ed. E. N. Zalta, http://plato.stanford.edu>), written by participants in the current debates, such as Christopher Hitchcock and Phil Dowe, touch on elements of RVC.

REFERENCES

ARMSTRONG, D.M. (1983). *What Is a Law of Nature?* Cambridge: Cambridge University Press.

BAUMGARTNER, M. (2008). 'Regularity Theories Reassessed', *Philosophia* 36: 327–54.

BEEBEE, H. (2006*a*). *Hume on Causation*. Abingdon: Routledge.

—— (2006*b*). 'Does Anything Hold the Universe Together? *Synthese* 149: 509–33.

BROWN, T. (1822). *Inquiry Into The Relation of Cause and Effect*. Andover, Mass.: Flagg & Gould.

—— (1851). *Lectures on the Philosophy of the Human Mind*. 19th edn. Edinburgh: Adam & Charles Black.

CLENDINNEN, J.F. (1992). 'Nomic Dependence and Causation', *Philosophy of Science* 59: 341–60.

COSTA, M.J. (1989). 'Hume and Causal Realism', *Australasian Journal of Philosophy* 67: 172–90.

DAVIDSON, D. (1967). 'Causal Relations', *Journal of Philosophy* 64: 691–703. Repr. E. Sosa and M. Tooley (eds.), *Causation*. Oxford: Oxford University Press 1993: 75–87.

—— (2005). 'Laws and Cause', in id., *Truth, Language, and History*. Philosophical

Essays 5. Oxford: Oxford University Press.

DRETSKE, F. (1977). 'Laws of Nature'. *Philosophy of Science* 44: 248–68.

DUCASSE, C.J. (1951). *Nature, Mind and Death*. LaSalle, Ill.: Open Court.

—— (1968). *Truth, Knowledge and Causation*. London: Routledge & Kegan Paul.

ELLIS, B. (2001). *Scientific Essentialism*. Cambridge: Cambridge University Press.

GOODMAN, N. (1983). *Fact, Fiction and Forecast*. 4th edn. Cambridge Mass.: Harvard University Press.

HAUSMAN, D.M. (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.

HEMPEL, C.G. (1965). *Aspects of Scientific Explanation*. New York: The Free Press.

HORWICH, P. (1987). *Asymmetries in Time*. Cambridge, Mass.: MIT.

HUME, D. ([1739] 1978). *A Treatise of Human Nature*, ed. L. A. Selby-Bigge and P. H. Nidditch. Oxford: Clarendon.

LEWIS, D.K. (1973*a*). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

—— (1973*b*). 'Causation', *Journal of Philosophy* 70: 556–67. Repr. Lewis 1986b: 159–213.

—— (1984). 'Putnam's Paradox', *Australasian Journal of Philosophy* 62: 221–36.

—— (1986*a*). *On the Plurality of Worlds*. Oxford: Blackwell.

—— (1986*b*). *Philosophical Papers II*. New York: Oxford University Press.

MACKIE, J.L. (1974). *The Cement of the Universe: A Study of Causation*. Oxford: Clarendon.

MILL, J.S. (1911). *A System of Logic: Ratiocinative and Inductive*. London: Longmans, Green.

MOLNAR, G. (2003). *Powers*. Oxford: Oxford University Press.

MUMFORD, S. (1998). *Dispositions*. Oxford: Clarendon.

—— (2004). *Laws in Nature*. London: Routledge.

PAP, A. (1952). 'Philosophical Analysis, Translation Schemas, and the Regularity Theory of Causation', *Journal of Philosophy* 49: 657–66.

PSILLOS, S. (2002). *Causation and Explanation*. Chesham: Acumen.

—— (2004) 'A Glimpse of the Secret Connexion: Harmonising Mechanisms with Counter-factuals', *Perspectives on Science* 12: 288–319.

—— (2006*a*). 'What do Powers do when they are not Manifested?', *Philosophy and Phenomenological Research* 72: 135–56.

—— (2006*b*). 'Critical Notice: Laws in Nature'. *Metascience* 15: 437–69.

—— (2007). 'What is Causation?', in Beena Choksi and Chitra Natarajan (eds.), *Episteme Reviews: Research Trends in Science, Technology and Mathematics Education*. Bangalore: Macmillan India.

—— (2008). 'Causal Pluralism', in Robrecht Vanderbeeken and Bart D'Hooghe (eds.), *Worldviews, Science and Us: Studies of Analytical Metaphysics: A Selection of Topics From a Methodological Perspective*. Singapore: World Scientific Publishers.

QUINE, W.V. (1969). 'Natural Kinds', in id., *Ontological Relativity and Other Essays*. New York: Columbia University Press.

RAMSEY, F.P. (1928) 'Universals of Law and of Fact', in D. H. Mellor (ed.), *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*. London: Routledge & Kegan Paul, 1978.

REID, T. (1788). *Essays on the Active Powers of Man*. Edinburgh: John Bell Smart.

STRAWSON, G. (1987). 'Realism and Causation', *Philosophical Quarterly* 37: 253–77.

—— (1989). *The Secret Connexion*. Oxford: Clarendon.

TOOLEY, M. (1977). 'The Nature of Laws', *Canadian Journal of Philosophy* 7: 667–98.

VENN, J. (1889). *The Principles of Empirical or Inductive Logic*. London: MacMillan.

WITTGENSTEIN, L. (1953). *Philosophical Investigations*. Oxford: Blackwell

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

WRIGHT, JOHN P. (1973). *The Sceptical Realism of David Hume*. Manchester: Manchester University Press.

CHAPTER 8
COUNTERFACTUAL THEORIES

L. A. PAUL

I hit the eight ball into the corner pocket. If I hadn't hit the eight ball, it wouldn't have rolled into the corner pocket. Many think that the counterfactual 'if I hadn't hit the eight ball, it wouldn't have rolled into the corner pocket' captures something essential and fundamental about this instance of causation: my hitting the eight ball caused it to roll into the pocket, since if I hadn't hit it, it wouldn't have rolled into the pocket. Counterfactual analyses of causation seek to exploit this insight by constructing analyses of the causal relation (or our concept of it) in terms of counterfactual dependence. The central thought behind a counterfactual analysis of causation is that the relation of counterfactual dependence between *E,* the *eight ball rolling into the corner pocket* and *C, my hitting the eight ball* somehow captures the fact that there is a causal relation between these events. That is, *C* causes *E* because the counterfactual 'if not *C*, then not *E*' is true. To the extent that this is successful, we have a counterfactual analysis of causation.[1]

Counterfactuals are subjunctive conditionals of the form, 'if it were the case that *A,* then it would be the case that *B*'. Counterfactual analyses of causation focus on counterfactuals that tell us what would have been the case if the world had been different. In the main, they focus on counterfactuals concerning temporally successive, suitably distinct events *C* and *E* that describe cases where, if *C* had not occurred, *E* would not have occurred. Some counterfactual analyses are developed in terms of probabilistic counterfactuals, for example, if *C* had not occurred, *E* would not have had the probability of occurring that it did have. For reasons of expediency, I will focus on analyses built with deterministic conditionals.

Such counterfactuals describe relations of counterfactual dependence. Many philosophers take counterfactual dependence between (successive, suitably distinct) events to be sufficient for causation (Lewis 1973*b*). If event *E* would not occur if event *C* were not to occur, then *C* is a cause of *E*. However, since there are intuitively clear cases of causation without simple counterfactual dependence, counterfactual dependence is not necessary for causation: it is not the case that *C* is a cause of *E only* when *E* depends on *C*. *E* might still occur, otherwise caused —even if *C* is a cause of *E* in the actual world (see the discussions of pre-emption and overdetermination in sects. 3.2 and 3.3 below). Hence, the causal relation between *C* and *E* cannot be reduced to the relation of counterfactual dependence between *C* and *E*.

David Lewis, the most prominent contemporary defender of a counterfactual analysis, tries to solve this wrinkle by taking the ancestral of the counterfactual dependence relation. If we call the relation of the ancestral of counterfactual dependence '*R*', *C* stands in *R* to *E* iff *E* counterfactually depends on *C,* or *C* is connected by stepwise counterfactual dependence to *E*.

Thus, even in cases where it is false that *E* depends directly on *C,* since if *C* had not occurred, *E* would have occurred (by being otherwise caused), as long as there is stepwise counterfactual dependence between *C* and *E, C* stands in *R* to *E.* A very simple and elegant account of causation could then take either dependence or stepwise dependence to be necessary for causation, and hold that *C* is a cause of *E* iff *C* stands in *R* to *E.* Call this the 'simple account'. David Lewis defended the simple account in his seminal (1973*b*) article, 'Causation' (reprinted in his 1986*a*). Defenders of similar accounts include Mackie (1965) and Lyon (1967).

Unfortunately, as Lewis himself quickly realized, the simple account is plagued with a host of problems. One obvious potential problem is that *R* is transitive, so if causation is the relation *R* then causation must also be transitive. Many, including Lewis, welcome this result, since they believe that the causal relation is transitive, but in recent years the transitivity of the causal relation has become controversial. (See Sartorio (2006) for an argument against the transitivity of causation, and Kvart (2001), Hall (2000), and Hitchcock (2001) for further discussion and debate.)

Many other problems have surfaced since Lewis's defence of the simple account, especially problems with varieties of pre-emption and overdetermination. As it turns out, the problems of pre-emption and overdetermination mean that the simple account is unworkable, and that a successful counterfactual analysis will need to do more than take the ancestral of dependence. Some of these issues are discussed in detail below, and the topic is dealt with at length in Hall and Paul (forthcoming). The many and varied problems with simple accounts have led to a proliferation of more complex counterfactual treatments of causation, for example, Ganeri, Noordhof, and Ramachandran (1996), Hitchcock (2008), Lewis (1986*c*; 2004*a*), McDermott (1995), Menzies (2004), Paul (2000), Schaffer (2005), Woodward (2003), and Yablo (2004).

1. METHODOLOGY

Counterfactual analyses have received a good deal of attention in recent years, resulting in a host of counterexamples and objections to the simple analysis and its descendants. The counterexamples are often complex and can seem baroque to the outsider (indeed, even to the insider), and it may be tempting to dismiss them as irrelevant or uninteresting. But while we may be able to ignore some counterexamples because the intuitions they evoke are unclear or misguided, the importance of investigating the causal relation via investigating counterexamples should not be underestimated. Consideration of specific cases is extremely important, for such cases can discover a previously unrecognized general issue for counterfactual analyses, and often give us insight into the nature of the causal relation itself, independently of any particular analysis of it, counterfactual or otherwise.

In particular, the close investigation of counterexamples involving cases of preemption (e.g. where *C* causes *E* and *D* does not, but if *C* had not occurred, *D* would have caused *E*) and overdetermination (e.g. *C* and *D* cause *E* and neither *C* nor *D* causes the other, and if *C* had not occurred, *D* would still have caused *E,* and if *D* had not occurred, *C* would still have caused *E*) can give insight into many of the deep and delicate issues surrounding an account of causation. The need for clearer and deeper accounts of the transitivity of causation, what a causal process

is, what it is to 'interrupt' a causal process, the role of omissions in causation, the role of the intrinsic character of the causal relation, and the connection between the intrinsicality of the causal relation and the character of the counterfactual dependence relation, was only recognized when counterexamples involving various sorts of preemption or overdetermination were discovered.

Apart from the need to investigate central counterexamples thoroughly, there are other methodological issues that affect investigations of causation and by extension any counterfactual treatment of causation. Recent work on causation has not been clear enough about the nature of the project or the problem that is to be solved. Is the project one of pure conceptual analysis, so that the problem is to develop an adequate analysis of a concept of causation? If so, is the concept of causation to be analysed a folk concept, a philosophical concept, a scientific concept, or something in between?

There are more variations that need to be sorted: conceptual analyses could be descriptive, focusing on describing or elucidating our actual concepts, or prescriptive, focusing on constructing an improved version of a concept of causation. Prescriptive treatments can combine an analysis of our ordinary concept with other desiderata to construct a philosophical concept of causation, draw on developments in science in order to challenge certain assumptions implicit or explicit in our ordinary epistemic stance towards the world, and learn from psychology to help us identify features of our causal concept. Ideally, a prescriptive account will combine all these techniques in order to eliminate internal contradictions and construct a clean, clear, and sensible concept of the causal relation that refines our unruly pretheoretical notions. Recent work in experimental philosophy draws on psychology for information about actual concepts: such work is obviously relevant to descriptive conceptual analyses, and may be relevant to prescriptive analyses as well (e.g. Hitchcock forthcoming, Hitchcock and Knobe forthcoming).

If the project is not conceptual analysis, is it ontological—a study of the ontology of the causation relation? Such a study could take the causal relation to be fundamental, or could take it to reduce to more fundamental entities. Further questions arise when we try to determine how we move from concepts to ontology (does a conceptual analysis give us an ontological analysis?) and what the role of our concept of causation is in determining the ontology of the relation. The best theories will probably mix some conceptual analysis with ontological analysis, if only to help clarify the subject under discussion. But what sort of a mix should we aim for?

Obviously, many people are interested in understanding how our concept of causation fits with the world, and to this extent we need to be clear on certain conceptual issues regarding causation even if our primary goal is ontological reduction. But the conceptual analysis of causation might also fit hand in hand with an examination of its ontology. Advocates who prefer to use conceptual analysis as a tool to carve out an ontological niche, Canberra-plan style (e.g. as in Jackson 1998), will defend a highly specialized account of conceptual analysis, usually one that is supposed to help us discover what is fundamental in the world via discovering what our words refer to. Such a plan for causation might proceed by analysing our concept of causation to define what the functional role of causation should be, and then taking the causal relation to be whatever relation (if any) in the world fills the role. One might also hold that conceptual analysis and ontological reduction are separate projects but necessarily

related, such that success or failure in one sort of approach may translate into success or failure in the other.[2]

A related approach might be to argue, following the structure of debates in the philosophy of consciousness, that conceptual analysis and related theses about conceivability and possibility can help us identify relevant metaphysically possible cases of causation.[3] One could even admit a role for conceptual analysis yet reject Canberra-plan approaches and other views with explicit requirements for how conceptual analyses are to be constructed and used. Such a rejection could still make room for some sort of conceptual analysis (which allows for attention to our philosophical intuitions about causation) to play an important role in helping us to distinguish between cases of causation and cases of non-causation. Conceptual analysis could then help to identify relevant features of the world needed to guide the development of a theory of causation and to target the subjects of the ontological reduction (see Hall 2006 for discussion of this idea).

Other positions about the interplay (or lack thereof) between conceptual analysis and ontological analysis can be staked out. One might wish to develop an analysis of a concept of causation while taking the ontology of any causal relation to be irreducible to any more fundamental relation. The goal of such an analysis would be to understand our concept or concepts of causation, independently of what the metaphysics of the relation (should such a relation even exist) would involve. At the other end of the spectrum, one might not be interested in conceptual analysis at all, seeking answers about causation from fundamental physics or other empirical sources. Some of those who reject conceptual analyses but who also refuse to make philosophy into a handmaiden to natural science might draw on robust metaphysical theories about the structure of reality and the nature of possibility as well as empirical sources to develop an ontology of causation.

Moving on, another methodological issue involves the role of normative or pragmatic factors in one's philosophical approach. A counterfactual-based approach could rely on pragmatics to construct an analysis of causation (for a very nice sample of such work, see Hitchcock 2008). Such an approach involves adopting a measure of pragmatism such that subjective or normative elements are built into the account. Pragmatic accounts have much to contribute to the literature on causation. One may rely on pragmatics to determine that a series of events or causal processes are relevant or irrelevant (or whether causal pathways are 'main' pathways or 'alternative' pathways) with respect to our causal judgements, to decide whether an effect should be included in the representation of the causal structure of a case, to establish the values of event variables to be assigned in causal models, or to determine truth values for certain counterfactuals.

However, these accounts face an objection: isn't such an account running together causal explanation with causation?[4] It seems right to say that causal explanation is mind- and description-dependent. It seems wrong to say that *causation* is mind- and description-dependent. Those who endorse a pragmatic account of causation will have to bite the bullet and deny this judgement. They may be willing to do so, given that pragmatic accounts will usually have a much easier time handling problems with pre-emption, overdetermination, and other serious difficulties for non-pragmatic reductive analyses. The ends may justify the means. But it is important to see that the metaphysician or philosopher of science who gives a pragmatic account of the causal relation or of the concept of causation is giving up on the

hardest part of the puzzle: she is giving up on the analysis of the causal relation as an objective, that is, a description- and norm-independent, relation in the world or of the causal concept as a description- and norm-independent concept of an objective relation. For many philosophers of causation, giving an account of causation or of our concept of causation requires the presumption that the relation and our concept of it are objective.

The benefit of giving up on the hardest part is that the defender of the pragmatic account can usually solve the worst of the knotty problems raised by the possibility of pre-emption, overdetermination, causation by omission, and the like, or at least make sense of such cases in an intuitively plausible way. The cost is that our natural presumption that we have an objective concept of causation and that causal relations between particular events are objective relations, must be given up. Those who prefer to avoid infecting causation with pragmatics (keeping the pragmatics for, e.g., accounts of *explanations* given in causal terms as opposed to accounts of *causation*) must continue to grapple with these exceedingly difficult counterexamples.

As I've indicated, then, the most ambitious accounts of causation are reductive. Reductive ontologies of the causal relation hold that facts about what causes what are fixed, somehow, by non-causal facts about what happens together with the facts about the fundamental (also non-causal) laws. Reductive conceptual analyses analyse our concept of causation in terms of other, more fundamental non-causal concepts. A reductive counterfactual analysis of the causal relation is an ontological analysis of how the causal relation reduces to ontologically more fundamental dependence relations, and a reductive counterfactual analysis of the concept of causation is an analysis of the concept in terms of (supposedly more fundamental) concepts of counterfactual dependence.

An analysis of the ontology of the causal relation could aim for a purely local reduction: to reduce the this-worldly causal relation to more fundamental this-worldly relations (e.g. Fair 1979; Dowe 2000), or might aim for a reduction that encompasses the actual world and a closely limited set of worlds very much like our own. A more ambitious reduction would be a general treatment of causation in terms of (reductive) supervenience: such an account would specify the supervenience base for the causal relation such that any possible world that has the supervenience base has causation.

Many metaphysicians find reductionism very appealing, although some explicitly reject it (e.g. Carroll (Ch. 13 below); Tooley 1987; 1990; and Woodward 2003). More recently, counterfactual-based accounts that partly reject conceptual reductionism have received a good deal of attention, particularly causal modelling accounts such as Pearl (2000), Halpern and Pearl (2005), and Hitchcock (2001). Causal modelling relies on antecedent notions of what caused what, or on how to assign values to variables in the model in order to represent causal patterns and causal manipulability. Again, to understand the logical space of theories of causation, it is important to see that these non-reductive conceptual analyses of causation engage in a different sort of—admittedly important—project from that engaged in by the metaphysician who is attempting to give a non-pragmatic, reductive analysis of causal concepts and (especially) from the metaphysician who is attempting to give a non-pragmatic, reductive analysis of causal relations.

The methodological precepts of reductive views need to be clarified before one can determine whether a particular reductive project need—or need not—be reductive across the

board. For example, if one's methodology requires everything in the world to supervene on patterns of contingently related instantiations of properties, then one's reductive theory of causation should include a reductive theory of laws of nature and an appropriately reductive account of modality. But a theory of causation can still be reductive without such broad reductionist commitments, for example, it could include a non-Humean account of laws or be primitivist about modality, as long as there were no stealthy assumptions about causation buried in these associated views. For example, any reductive ontological analysis of the causal relation will require the laws of nature that serve as part of the supervenience base for causation in a world to be specifiable in non-causal terms, even if such laws are treated in non-Humean terms. Laws making explicitly causal claims such as 'x's having P causes y's having Q' (e.g. as we see in Tooley 1990: 226) cannot be fundamental in a reductive treatment of causation: the acceptable law must be stated as 'if x has P, y has Q'.

No matter what method one applies to the analysis of causation, one must be clear about the role of thought experiments, intuitions, and the relevance of various conceptual and metaphysical possibilities. The rich history of conceptual analysis makes the role of such philosophical tools reasonably clear, even if issues such as the relation between conceivability and possibility remain controversial. But the difference between the goals and presuppositions of conceptual analysis and those of unabashed ontological analysis mean that the precise role of thought experiments, intuitions, and counterexamples could also be quite different in determining the success or failure of an account of causation. In other words, exactly what sort of project one is undertaking may influence one's approach to dealing with challenges to various proposed analyses, especially with respect to how a proposed counterexample needs to be addressed (this point is often lost among the thicket of the various analyses and objections proposed). For example, if an account of causation only aims to give an account of the causal relation in the actual world (and worlds with the same physical laws), then counterexamples involving bizarre laws of magic at distant possible worlds are irrelevant. Related issues crop up when deciding whether one is tackling the hard problem of causation or whether one is content to accept non-reductive or pragmatic elements as part of one's account. Philosophers working on ontological analyses (or analyses involving a mix of ontological and conceptual analysis) have not been precise enough about how the widespread rejection of pure conceptual analysis requires a change of methodological perspective.

A related problem for those working on the ontology of the causal relation concerns the breadth of their focus on ontology: are they content to rely on results derived from conceptual analyses for associated topics, such as the ontology of modality or of laws? If they reject, in principle, the project of constructing conceptual analyses as a tool for guiding reductions, then they should re-examine their commitments to related topics which may have been influenced by considerations motivated by conceptual analysis. For example, one might wish to revisit the semantics of counterfactuals or the theory of laws one endorses, since, for example, accepting more primitivism in one's accounts of what laws are or in determining the direction of counterfactual dependence can simplify or change the task for the ontologist of causation. If, instead, one's view is that conceptual analyses are appropriate guides for some reductions but not others, a consistent accounting of which topics are suitable for conceptual analysis— and why—is called for. The interaction between all of these methodological issues and accounts of causation (counterfactual or otherwise) is an area where further research is

desperately needed.[5]

In what follows, I will focus on reductive, objective counterfactual analyses of causation and will try to be clear about the way that various methodological presuppositions infect our assessments of the strategies and problems for counter-factual analyses. I will distinguish conceptual and ontological variants of these analyses of causation only when it is absolutely necessary.

2. Motivation

Why attempt to develop a reductive conceptual analysis of causation or a reductive ontological analysis of causation? A general theoretical motivation for a reductive analysis of causation is that such an analysis would be deeply related to many other central philosophical topics, and would serve as a tool for philosophers, scientists, and others to use, the better to understand analyses of laws of nature, events, properties, practical reasoning, objects, probability, determinism, perception, mental causation, the existence of God, reference to and knowledge of the external world, agency, free will, moral responsibility, and legal responsibility.

Why attempt to develop a *counterfactual* reductive analysis of causation? One reason might be that there is no obviously superior alternative available. But a better motivation for adopting a counterfactual analysis starts with the intuition that cases of counterfactual dependence are cases of causal dependence. There seem to be systematic connections between our judgements that Es depend counter-factually on Cs and our judgements that Es are caused b y Cs, and many theorists hold that the presence of counterfactual dependence (or its ancestral) is clearly sufficient for token causation. This is a connection between counterfactual dependence and the notion of influence and manipulability: if an event depends on another event, it is influenced or can be manipulated by that event. In everyday life as well as in the empirical and social sciences, causes are identified by the determination of manipulation: Cs are causes of Es if changing Cs changes the Es, that is, if we can manipulate Es by manipulating Cs. In this way, experimental settings are designed to test for the presence of causation by testing for the presence of counterfactual dependence. (For discussion, see e.g. Winship and Morgan 1999.) The idea here is that the counterfactual dependence of E on C points to an underlying causal mechanism operating between C and E, and so detecting dependence is our best way reliably to infer the existence of the causal mechanism. (It is worth noting that Pearl 2000, Woodward 2003, and others have developed causally non-reductive interventionist accounts of causal modelling based on this idea.)

Relatedly, connecting causation with counterfactual dependence permits a 'black box' strategy that seems to be absolutely essential for our epistemic access to causation. If all we know about E is that it counterfactually depends on C, this is enough to infer that E is caused b y C, even if we don't know anything else about the relation between C and E.[6] This is important for our everyday navigation of the world. If we had to know, for example, whether there was a process or a transfer of energy in order to know that there was causation, it would be much harder to recognize the causal relation. Since we seem to be able to determine the presence of causation just on the basis of dependence or manipulability, this gives us a strong

prima facie case for the idea that mere counterfactual dependence is sufficient for causation, and perhaps even for the idea that the causal relation is reducible to a (suitably qualified kind of) counterfactual relation.

A counterfactual analysis also has a reasonable amount of flexibility in responding to test cases, which is essential in order to treat crucial examples adequately. Lewis (1973*b*) argues that the counterfactual analysis, when combined with a sophisticated approach towards the semantics of possible worlds, correctly solves problems with common causes. Common cause cases are where *C* causes *B* and also causes *D* which, slightly later, causes *E*. *E* occurs later than *B*.

Consider the 'neuron' diagram of causation in Fig. 8.1. Capital letters name events. A filled circle represents the occurrence of an event, an empty circle represents the absence of an event, a line with a triangular head represents a causing of an event (a line with a circular head, not shown, represents the inhibiting of or the prevention of an event). Temporal progression from earlier to later is represented by reading from left to right.

One might think that *E* counterfactually depends on *B*, since had *B* not occurred, this would mean that *C* had not occurred, hence that *E* would not have occurred either. Not so, says Lewis: 'had *B* not occurred, *C* would not have occurred' is a *backtracking* counterfactual, counterfactuals that involve subjunctive conditionals where the time of the event described in the consequent precedes the time of the event described in the antecedent. Backtrackers are barred from use in evaluations of counterfactual dependence in worlds like our own. (Worlds with very different laws from ours might allow backtrackers. But in worlds with laws like ours, many backtrackers imply a violation of our laws, or at least of our laws plus certain assumptions about initial conditions.) Thus, even if *B* had not occurred, *C* would have occurred anyway, and *E* would have occurred, so *E* does not depend on *B*. If appropriate restrictions on backtrackers can be upheld, the counterfactual analysis rightly tells us that *C* is a cause of *E* but that *B* is not a cause of *E*.
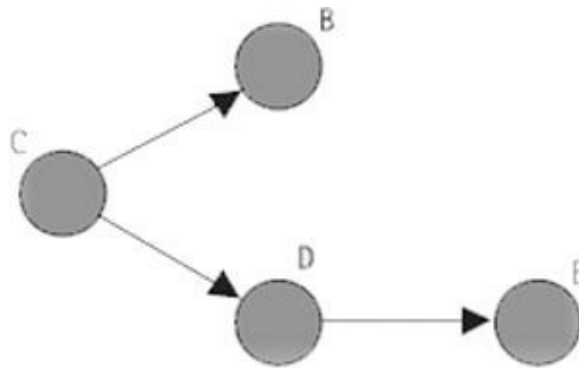
Fig. 8.1
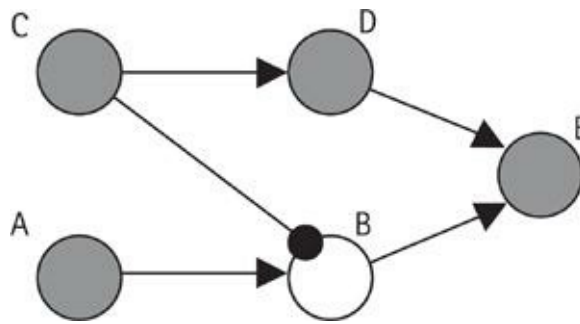
Fig. 8.2

A counterfactual analysis of causation can also handle certain sorts of cases of causal pre-emption; cases where events compete to cause an effect, but one or more of the competitors are pre-empted, in other words, the competitor does not succeed in causing the effect. Consider Fig. 8.2, which represents a standard case of causal pre-emption.

In Fig. 8.2, A and C are competing to cause E. C causes E while inhibiting B, thus pre-empting A (and B) from causing E. (The line that ends in a dark circle represents an inhibitory stimulus.) Assume that E would have occurred with exactly the same intrinsic characteristics if it had been caused by A. This case creates problems for simple counterfactual accounts, since E does not counter-factually depend on C.

Lewis's counterfactual analysis has a neat treatment of the case in Fig. 8.2. Since, on the assumption that causation is transitive, we take causal dependence to be the ancestral of counterfactual dependence, we see that C is a cause of E because there is stepwise counterfactual dependence between C and E.[7]

Another important motivation for a counterfactual analysis is that it can support causal claims in situations where there is no process, transfer of energy, or series of events between cause and effect. Such situations involve *negative* causation, that is, causation involving omissions or absences. For example, my failing to set my alarm causes me to miss my class, and my failure to hire a landscaper causes me to feel confused about what to do with my garden. The absence of a landscaper can cause my garden to run amok, and my inability to fix the garden can prevent my house from being sold. To the extent that these examples involve omissions or preventions of events, we can call them examples of negative causation.

There are deep and puzzling issues surrounding causation by omission, since omissions, by definition, are not existents of any sort—they are *absences* of existences. Lewis (2004*b*) draws the striking conclusion that if omissions can be causes and effects, then there can be causation without a causal relation (because a relation requires relata, and at least one relatum goes missing if an omission is involved). For this reason, omissions create especially difficult problems for conceptual analyses of causation that want to provide analyses of causation that do not separate off omissions as special cases.

Ways around some of the problems with omissions might involve translating claims about omissions into claims about certain regions of spacetime (although it can be difficult to develop a satisfactory account of which region or regions are being picked out, and to cover all cases, both possible and actual) or by defending fact causation, where the causal relation can

take linguistic entities such as negative facts or sentences as relata. (For defences of fact causation, see McGrath (forthcoming); Mellor 1995; 2004.) Another issue involves the need to think in terms of types: it is usual to hold that an omission 'occurs' iff no event of a certain type occurs. But giving an adequate specification of such types is no easy matter. (One might think, in addition, that there is a normative component to the specification. See Beebee (2004) for critical discussion and McGrath (2005) for a defence of this idea.) These issues are pressing, serious problems for any analysis of causation, and much more research needs to be done to address the deep puzzles that causation by omission raises. But once we set these general problems with omissions for all analyses of causation aside, the important point that remains is that a counterfactual analysis of causation does a much better job of handling causation by omission than many of its rivals.

The accounts most threatened by causation by omission are those relying on transference, processes, or sufficiency based on actual events that instantiate properties of fundamental laws. Consider Fig. 8.3.

Take C to be the falling of a large boulder, and A to be Suzy's pushing Billy out of the way. E is the crushing of Billy under the boulder. The occurrence of A prevents the occurrence of E: Suzy saved Billy's life by preventing the crushing, that is, by causing an omission.

This creates problems for any account of causation that requires there to be some sort of process or transference to link the causal relata or that cannot accommodate absences as causal relata (at least as relata in some sense). The problem particularly affects theories that reduce causation to processes or to sufficiency under fundamental laws, since unless there is a reconstruction of the causal sequence that involves an alternative process or a fundamental law, they must hold that no causation occurs.[8] (Dowe (2000: ch. 6) and Fair (1979) treat such cases in terms of different counterfactuals about process causation: it is unclear how they would handle Lewis's (2004b) example of the deadly void. For an alternative assessment of what is going on, see Beebee (2004).)
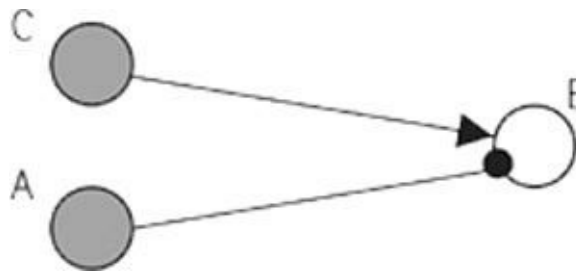
Fig. 8.3

Related cases such as the one in Fig. 8.4 involve what Ned Hall (2004) calls *double prevention*.

Such cases involve chains of (possible or actual) preventings (usually two in a row, hence 'double' prevention). A standard case involves Billy, who throws a rock at a bottle, and Suzy, who intends to prevent him by stopping his arm. As Suzy (B) reaches out to stop Billy's throw, Hillary reaches out (C) and prevents (D) Suzy from reaching towards Billy. Billy throws the rock (E). Hillary's grab prevents Suzy's block from preventing Billy's throw. In this case, Hillary's act is among the causes of Billy's throw, even though, as Hall points out, there is no

process or fundamental regularity instantiated between Hillary's act and Billy's throw. Double prevention, once it is recognized, seems to be everywhere: guns firing, physiological processes, and everyday activities all involve double prevention.[9] It is simply unacceptable for an account of causation to hold that sequences involving double prevention are not causal.
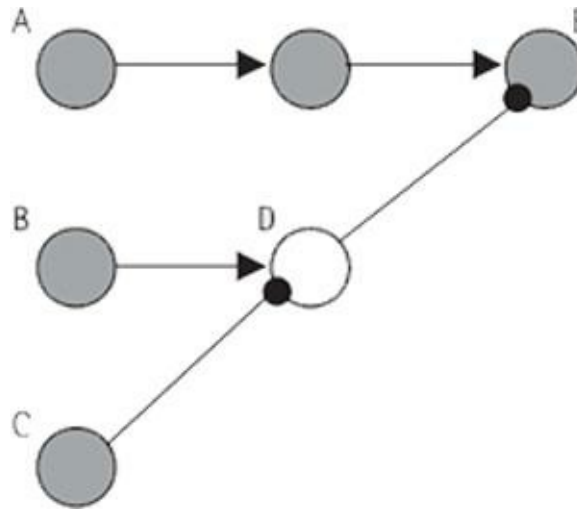
Fig. 8.4

Because cases like those in and exhibit the requisite counterfactual dependence, the counterfactual theorist can hold, correctly, that such cases exhibit causation. If Suzy hadn't pushed Billy out of the way, he'd have been crushed. If Hillary hadn't grabbed Suzy's arm, Billy would not have thrown the rock. A counterfactual account makes such causation easy—all that is required is a certain sort of dependence—and thus can get the right answer when we have causation by omission. If, on the other hand, one requires that there be a transfer of energy or momentum between cause and effect for causation or denies that omissions can be causes or effects, cases of prevention and of double prevention don't count as cases of causation without a good deal more work.

3. PROBLEMS

Although the elegance, relative simplicity, flexibility, and intuitive power of a counterfactual analysis are strong arguments in its favour, many problems have been raised in the literature. Some apply to counterfactual approaches generally, such as worries about circularity or the order of ontological dependence. (Which is more fundamental, causation or counterfactual dependence?) Others involve the way particular analyses deliver verdicts on cases that are at odds with our intuitions.

For reasons of space, I will consider only a few of the most central problems facing counterfactual analyses of causal concepts and the causal relation, and I'll look at this small sample in a fairly selective way. I'll discuss the general problem of circularity, and then consider the more specific and interrelated problems of preemption and overdetermination. As a result, I will set aside many problems and puzzles, including those involving indeterministic

causation, distinctions between causation and causal explanation, trumping, the nature of the causal relata, further discussion of causation by omission, and transitivity. A close examination of the interrelated nature of pre-emption and overdetermination will help the reader to grasp some of the deepest and most central challenges for counterfactual analyses of causation: an extended treatment of many of the issues elided in this chapter can be found in Hall and Paul (forthcoming).

3.1 Circularity

An account of causation based on counterfactual dependence requires a semantics of the counterfactual relation, that is, an interpretation of counterfactual claims that will determine the dependence relation that such claims rely upon. If the interpretation of counterfactual claims is determined by assessing causal claims, there seems to be a problem with circularity: one's account of the causal relation in terms of counterfactual dependence requires an account of counterfactual dependence in terms of causation. Perhaps such an account could still be informative: as long as we have enough conceptual access to causation to break into the circle, we can gain important insights by developing the tight connections between dependence and causation. (Woodward (2003) defends the view that a circular analysis is still informative.)

Lewis (1979) does seem to be employing causal notions to guide the development of his semantics of dependence, but such notions are used merely as rules of thumb to guide the development of the account of dependence in terms of qualitative similarities (of facts and laws) between worlds. The ultimate semantics is one that uses qualitative similarities as the ontological basis for evaluations of counterfactual dependence, and hence is not circular. (For details, see the discussion of the semantics of counterfactuals above.) Lewis's account is not without its problems, however, for as Elga (2001) argues, Lewis's account of the truthmakers for counterfactuals in our world is flawed. Lewis holds that certain asymmetries of overdetermination allow us to exclude backtracking counterfactuals in a systematic way, which is important both for an adequate account of counterfactual claims in our world and for an adequate counterfactual analysis of causation for worlds like ours where there is no backwards causation. The problem is that, as Elga points out, Lewis's interpretation of how our world exhibits certain asymmetries of over-determination conflicts with implications of our fundamental dynamical laws. As a result, depending on what sort of reduction or analysis one wishes to carry out, the semantics of counterfactuals and its attendant counterfactual analysis of causation needs to be patched up. In particular, advocates of a Lewis-style conceptual analysis of causation need to patch up the semantics for counterfactuals in a way that leaves it flexible enough to exclude backtracking conditionals for worlds like ours while still accommodating the possibility of having worlds in which there is backwards causation. The options and issues here are canvassed in Collins, Hall, and Paul (2004) and discussed in detail by Price and Weslake (Chapter 20 below).

3.2 Pre-emption

Cases of pre-emption are cases where C causes E, but if C had not caused E, one or more

back-up causes (merely potential causes) would have caused *E* instead. Such cases are cases of pre-emption (as opposed to cases of overdetermination—see sect. 3.3 below) because *C* is a cause of *E* while the back-up causes are merely *potential* causes. We have already considered a version of so-called *early preemption* with
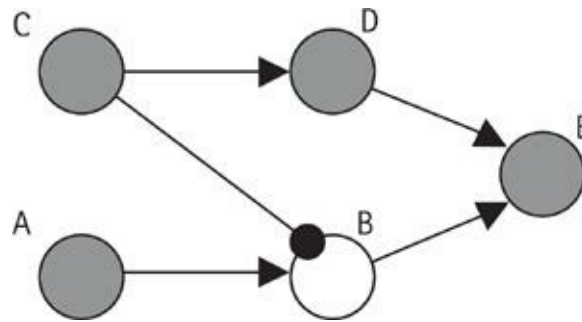
Fig. 8.2.

The pre-emption is called 'early' because the interruption or modification of the back-up processes occurs before *E* occurs. A counterfactual analysis without bells and whistles tells us that *C* is a cause of *E* iff, if *C* had not occurred, *E* would not have occurred. But here is a case where if *C* had not occurred, *E* would have occurred anyway, albeit otherwise caused. Yet *C* is a cause of *E*.

The case brings out a deep problem for a counterfactual account: counterfactual dependence is sensitive to extrinsic factors such as extraneous events that function as possible causes or back-ups. In other words, the relation of counterfactual dependence between cause and effect can be affected by non-causal goings-on in the neighbourhood. The problem is that we are inclined to judge that causation should not be sensitive to these sorts of extrinsic factors: intuitively, when *C* causes *E*, it does so whether or not there are other events around, assuming that these other events are not causally or otherwise necessarily connected to *C* or *E*. This is a way of saying that whether or not *C* causes *E* is independent of other entities not causally or otherwise connected to them (apart from the laws).

Yet we are trying to analyse causation in terms of counterfactual dependence. If so, how can such an analysis be correct? The most persuasive mitigating factor with regard to this problem is the competing intuition that counterfactual dependence seems to capture something deep and necessary about causation. If so, perhaps the answer is that an analysis of causation needs to have counterfactual dependence as a central part of it, even if the final analysis involves extra components designed to insulate (if such can be done) tests for dependence from extrinsic noise. While such an approach seems reasonable, counterfactual accounts have, so far, failed to make it work, making it unclear how to resolve these competing issues. (Though see Hall 2004 and Hall and Paul forthcoming for various suggestions about how to resolve the conflict.) This makes trouble no matter what sort of counterfactual analysis we are trying to develop, since a conceptual analysis of the concept of causation as a concept of counterfactual dependence would then seem to have mistargeted the relevant concept, and an ontological

analysis would seem to have misidentified the reduction base.

Setting these deeper concerns aside, how do we address early pre-emption? In Fig. 8.2, the modification of the back-up process is an outright prevention: C interrupts the causal process between A and E by preventing B. However, cases of early pre-emption can involve other ways of modifying the causal process besides preventing an event. Perhaps, when C occurs, C modifies B in such a way that B still occurs, but would no longer be able to cause E. If so, then when C occurs, C is a cause of E and A and B are not. Yet, if C had not occurred, E would still have occurred, caused by A and B. See Hall and Paul (forthcoming: sect. 5) for discussion of just such a case.

One seemingly obvious quick fix won't do the job. You might think that whatever A and B cause, it isn't E, since E is something that is necessarily caused by C. This solves the puzzle by individuating E by its causes. But recall that we are trying to give an account of what causation *is*. We cannot individuate events by their causes and effects while reducing or analysing causation in terms of counterfactual relations between events, since without an independent account of which events occur, there is no way to determine which counterfactual relations exist!

Counterfactual theories of causation usually solve the problem of early preemption by relying on the transitivity of the causal relation. Instead of an analysis where C causes E iff, if C had not occurred, E would not have occurred, take causation to be the ancestral of the counterfactual relation, that is, take C to be a cause of E iff, if C is connected to E by a series of events $D_1, D_2, \ldots D_n$ such that if C had not occurred, D_1 would not have occurred, and if D_1 had not occurred, D_2 would not have occurred ... and if D_n had not occurred, E would not have occurred. This modified counterfactual analysis, when applied to the case in Fig. 8.2, takes C to be a cause of E because E depends on D, and D depends on C. So C is connected to E by a chain of counterfactual dependencies, even if E does not depend on C outright. (This solution only works if backtracking in relevant cases is prohibited: see Lewis (1973*b*) for discussion.)

Another sort of pre-emption that has received a lot of attention in the literature is *late pre-emption*. Late pre-emption is best described as pre-emption where C causes E, but pre-empted back-up processes are not interrupted until E occurs.[10] A textbook case of late pre-emption involves a pre-empted back-up causal process that is interrupted because the pre-empting cause brings about the effect before the back-up cause can (under the laws). Such a case is represented by Fig. 8.5.
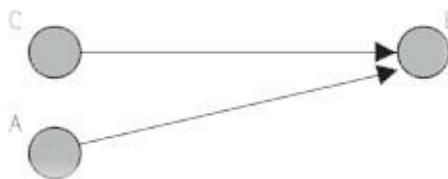
Fig. 8.5

Here, the pre-empting cause, C, causes E just before A would have caused E (this is

represented by the arrow from *A* failing to extend all the way to *E* at the time *E* is caused).^{11}

As with early pre-emption, the trouble is that *C* is a cause of *E*, but *E* does not depend on *C*: had *C* not occurred, *E* would still have occurred, since *A* would have caused it. But this trouble cannot be fixed by an appeal to transitivity, since whether or not there are events in the causal chain between *C* and *E*, there is no point at which *E* counterfactually depends on any such intermediate event (as it did on *D* in Fig. 8.2, above, in our sample case of early pre-emption). Thus we cannot take *C* to be a cause in virtue of being connected to *E* by a chain of dependencies.

Two quick fixes come to mind. First, one might try to individuate effects by their causes, so that *E* couldn't be caused by *A* after all. As I noted above, this approach cannot be used to solve pre-emption puzzles, since such puzzles arise within an attempt to develop a reductive analysis of causation. But there is a related move that might seem more promising: take events to be fragile, that is, to have extremely well-defined essences. I've already noted that in a textbook case of late preemption, *A* can't cause *E* when *C* causes *E*—if *A* had caused *E*, it would have caused *E* to occur a little bit later than it actually did. The fragility strategy exploits this fact by taking events to be temporally fragile, that is, taking events to be such that they could not have occurred any earlier or later than they actually did. If so, then in Fig. 8.5, *E* depends on *C*, since had *C* not occurred, *E* would not have occurred. Lombard (1986) and Coady (2004) develop fragility approaches.

Many authors, most notably Lewis (e.g. 1986*b*; 1986*c*; 2004*a*), have been sceptical of the fragility solution, denying that events must have fragile temporal essences. If events are temporally fragile, then when that event occurs, it is impossible for it to have occurred earlier or later than it actually did. Note that simply accepting that temporally fragile events exist isn't enough to solve the problem, since it seems plausible, of course, that some temporally fragile events exist. The solution needed to solve the late pre-emption problem is not the thesis that there exist temporally fragile events, but rather that *every* event is fragile with respect to the time it is located at. Otherwise, it is easy to design counterexamples involving the preemption of one of the non-temporally fragile events.

There is, however, a more plausible alternative: instead of taking events to be temporally fragile, take causal counterfactuals to be fragile. (See Paul 1998*b*.) On this approach, we exploit the fact that if *C* had not caused *E*, it would have occurred later than it actually did, but without requiring a special metaphysics of *E*. Instead, we define causation such that (along with the usual caveats), had *C* not occurred, *E* would not have occurred *when it actually did*. This requirement correctly classifies *C* as a cause. However, the solution seems to be arbitrarily specific to times, and accordingly, Lewis (2000; 2004*a*) and Paul (2000) extend it to include counterfactuals that are sensitive to a wide range of characteristics, such that, had *C* not occurred just as it actually did, *E* would not have occurred just as it actually did. The final version of the view defended by Lewis is the thesis that, if whether, when, and how *C* occurs influences *to a suitable degree* whether, when, and how *E* occurs, *C* is a cause of *E*.

These proposals, while capturing a part of the content of our concept of causation, have been criticized. Although the influence of one event on another seems to be a significant marker for causation since it is the primary empirical tool for the determination of causation, it can lead one into error. Some object that using cases that purport to show that changing just any characteristic of an event is not enough to merit causal status. (Paul (2000) defends this

result, given that causes and effects are property instances rather than entire events, while Lewis (2000; 2004*a*) denies it, adding the rider that the change in characteristics must be enough of a change to count as causation.) Others object that there are cases where *C* is not suitably influential in bringing about *E*, yet intuitively, *C* is a cause of *E*. A case of late pre-emption where the back-up process would have caused the effect to occur with the very same properties as *C* caused it to have, *modulo* reasonably plausible assumptions about what would have occurred if conditions were different, would be an example. These cases do seem to create problems for the influence theory. (For critical discussion of the influence theory, see Hall and Paul forthcoming; Schaffer 2001; Strevens 2003.)

Two especially stubborn kinds of cases of late pre-emption—call them *esoteric* cases of late pre-emption—for any sort of counterfactual analysis that is supposed to be a conceptual analysis of causation involve (1) cases involving action at a distance and (2) versions of cases discovered by Goosens (1979) involving multiple or infinitely many pre-empted alternatives. Lewis (1986*c*) considers instances of (1) but dismisses them as so far-fetched that his analysis need not address them. Given that Lewis is developing a conceptual analysis as opposed to an ontological reduction of causation, it would be helpful to have additional supporting arguments for this claim.

Even if we have a principled reason for setting aside cases involving action at a distance, esoteric cases of sort (2)—Goosens' cases—create trouble. To understand this sort of case, consider again the simple pre-emption diagram in <u>Fig. 8.2</u>. The causal chain started by *C* causes *E*, while the causal chain started by *A* is preempted. A Goosens-style case builds on this simple case by having many or infinitely many alternative back-up causal chains, each one being pre-empted just after the next. Instead of a single back-up, *A*, there are infinitely many (or as many as are needed, if time is discrete) backups $A_1, A_2, ... A_n$. The trouble here is that at every time during the causal process from *C* to *E*, there exists an uninterrupted back-up process. Such cases create problems for the transitivity solution to pre-emption since there is no point in the process from *C* to *E* where *E* depends on an intermediate event that depends on *C*. A third sort of esoteric late pre-emption is discussed by Paul (1998*a*) and Lewis (2004*a*) where there is no action at a distance, but where one cause pre-empts another without preventing events in the back-up process.

Strangely, Lewis (1986*c*) also dismisses Goosens-style cases as 'too far-fetched' and thus irrelevant to the success of his analysis. But it is unclear how such a verdict could be justified when one is pursuing a conceptual analysis: there seems to be no conceptual incoherence in the case, and so no reason to dismiss it as irrelevant. (After all, if cases involving magic and the like are acceptable, as Lewis (2004*a*) seems to think, then why not a Goosens case? Lewis (2004*a*) even accepts action-at-a-distance cases, effectively reversing his (1986*c*) view about cases of type (1).) We should take such cases seriously. A Goosens-style case where *E* does not depend on *C* because of the existence of back-up processes such that had any of those back-up processes caused *E*, *E* would have occurred exactly when and how it did occur, creates trouble for a broad range of conceptual analyses of causation in terms of counterfactuals.

3.3 Overdetermination

Causal overdetermination in its standard form involves symmetrical causal contributions

by multiple causal processes. Intuitively—but only very roughly—speaking, overdetermination can occur when more than one event, where each such event is part of a distinct, sufficient causal process, causes an effect. Fig. 8.6 gives a simple representation of a classic case of overdetermination.
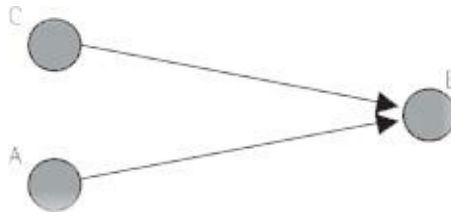
Fig. 8.6

In this case, *C*'s firing is sufficient to cause *E*'s firing in just the way it actually did. *A*'s firing is also sufficient to cause *E*'s firing in just the way it actually did. *C* causes *E*'s firing and *A* causes *E*'s firing. Although the intuitive idea of overdetermination seems simple enough, being precise about the deeper idea it involves, as well as getting the exact definition of causal overdetermination right, is extremely difficult.

To see why, start by considering the usual sort of case used to characterize overdetermination: two rocks, one thrown by Billy and one thrown by Suzy, hit a window at exactly the same time, shattering it. If this is a case of overdetermination, each rock-throwing alone causes the shattering. On a theory of causation where events are individuated robustly, that is, such that the very same shattering would have occurred whether there was one rock or two, this counts as a case of overdetermination.

But notice that if one counts the shattering by two rocks to be a numerically different event from the shattering by only one rock, and the world is such that if Billy hadn't thrown the properties of the shattering would have been different, and such that if Suzy hadn't thrown the properties of the shattering would have been different, then this is really a case of joint causation instead of overdetermination. (Billy and Suzy jointly cause the shattering, they don't overdetermine it.) For this reason, while the case of Billy and Suzy shattering the window can be treated as overdetermination, it is uninteresting to do so.

The interesting problems arise when we try to move beyond the sort of over-determination we get when events are individuated robustly. For overdetermination with fine-grained events, where a difference in properties amounts to a difference in events, we need to describe an example such that that the shattering occurs *precisely* the same way, whether one rock shatters it or two. Call this *fine-grained overdetermination*. But surely, if the Suzy and Billy case is supposed to occur in (a deterministic version of) a world like our own, such a case is physically impossible. For example, each rock contributes a certain amount of force to the shattering of the window, and under the laws, the properties of the shattering will be affected by an increase or decrease in the force of the impact. So, assuming determinism, claiming that the shattering must be exactly the same whether there are two rocks or one requires a significant departure from (deterministic versions of) the laws of our world. (Paul 2007 calls this the problem of *additivity*.)[12]

Unless there are clean cases like those described in Fig. 8.6 above, where, for example, the firing of *E* is stipulated to occur when a certain activation threshold is met, and where *A* and *C* each contribute, in precisely the same way and at precisely the same time, sufficient activation energy for *E* to fire, and where neither *A* nor *C* is pre-empting the other nor jointly causing *E*, fine-grained overdetermination does not seem to be physically possible (unless we can sort out a way to avoid the multiple subsumptions under the laws that additivity involves). For the example to work, don't think of *A* and *C* as each contributing a part of the energy needed to cause *E*'s firing—that way lies joint causation. Instead, by stipulation, *A* causes *E* in exactly the way it does in Fig. 8.6 even if *C* is absent, and *C* causes *E* in exactly the way it does in Fig. 8.6 even if *A* is absent.

In the end, such a clean case may not be physically possible. If so, this has implications for certain arguments about overdetermination and the actual world, especially when we consider debates about the possibility of mental or 'higher level' causation. On the assumption that token mental properties, events, or agents are not ontologically reducible to something more fundamentally physical, mental causation, agent causation, and other sorts of higher-level causation seem to involve fine-grained overdetermination. If so, and if such overdetermination is physically impossible, then these sorts of higher-level causation are physically impossible. (See Paul 2007 for further discussion.)

For the purposes of developing an analysis of causation, we can set this worry aside for now. Surely it is metaphysically possible to have a clean case of over-determination, and, given a few adjustments to the laws, it is perhaps even physically almost-possible. If so, this sort of case needs to be addressed by any counterfactual account that takes itself to apply more broadly than to worlds with physical laws exactly like our own.

The deeper issue underlying fine-grained overdetermination becomes clearer when we consider a clean case: how, exactly, can *A* and *C* each cause *E* if they are not causing it jointly? What work is being done by the much-needed stipulation that each cause brings about the effect just as it would if the other cause were absent? A way to express the worry is that it seems as though fine-grained over-determination requires too much causation.

Note how differently we feel about the clarity of cases of fine-grained over-determination versus that of cases of early and late pre-emption. In the preemption cases, the way each event makes (or doesn't make) its causal contribution is perfectly clear. But in fine-grained overdetermination cases, it just isn't clear how each cause is bringing about the effect all on its own, given that another cause is also bringing about the effect all on its own and the causation is not joint causation.

This is not an objection that assumes there is a certain amount of 'causal fluid' in the world that can bring about an effect and takes overdetermination to violate some sort of principle of the conservation of such fluid (see Sider 2003). Rather, it is the suspicion that there is a deep conceptual puzzle here about how the concept of overdetermination can fit with the concept of a sufficient cause as something that is entirely responsible for bringing about an effect. Given that *C* causes *E*, simply saying that *A* also causes *E* does not explain how *A* makes its causal contribution to the production of *E*. It merely reiterates that *A* *is* a cause of *E*. The worry is that fine-grained overdetermination seems to violate the natural intuition that—at least with cases that do not involve omissions—the complete causal character of a causal chain is fixed solely by its intrinsic character plus the laws.[13]

Admittedly, given the variety of ways events can cause other events, intuitions about the role of intrinsicality in a causal analysis are controversial. But this controversy is mitigated to some extent once we see how, in a simple case like the one in Fig. 8.7, causal character is fixed by the laws and its intrinsic character.

Fig. 8.7

In Fig. 8.7, how can it be the case that the causal facts about E can be changed without changing the intrinsic causal character of the causal chain between C and E? (The causal facts about E are changed because when we add A as a cause to get the case depicted in Fig. 8.6, E is overdetermined.) This simple case seems to show that intrinsicality matters.

Controversy about whether causal character is fixed solely by the laws and intrinsic character aside, the existence of the conceptual puzzle means that it is unclear, intuitively speaking, how fine-grained overdetermination works. Fans of counterfactual analyses try to exploit this lack of intuitive clarity, since counterfactual accounts have particular difficulties with cases of overdetermination. The main problem for counterfactual analyses derives from the sensitivity of counter-factual dependence to factors such as the presence of an overdetermining cause. E does not counterfactually depend on C, because A also causes it. Likewise, E does not counterfactually depend on A, because C also causes it. The defender of counterfactual analyses seems to be forced to fall back on one of two options: neither C or A caused E, for E is counterfactually dependent on neither of them, or the mereological sum of C and A caused E, for E is counterfactually dependent on this sum.

Holding that neither C nor A is a cause of E is unconvincing. E was caused, and C and A each seem to have caused it. How can it make sense to say that neither C nor A is a cause of E? The latter option is more appealing: the *sum* of C and A caused E. But note that this won't get the counterfactual analyst as much as might seem. Taking the mereological sum of C and A as the cause does not mean that we are taking C as a cause and A as a cause: instead, the mereological sum CA is a cause while neither event alone is.

This result is bizarre. How can the mereological sum of A and C be a cause of E while neither A nor C alone is a cause, joint or otherwise? It is unclear why the presence of C makes A unable to be a cause without being a member of the mereological sum of A and C, and likewise unclear why the presence of A makes C unable to be a cause without being a member of the mereological sum of A and C. Counterfactual accounts of causation violate the natural intuition that it seems right to say that if E is overdetermined then A is a cause of E and C is a cause of E, whether or not A and C compose a mereological sum.

For the counterfactual analyst who wishes to construct a reductive definition of overdetermination, the challenge is difficult. Further problems arise with the mereological sum treatment of overdetermination when we see that this is not enough to distinguish cases of overdetermination from joint causation, trumping (see Schaffer 2000*b*), or even early and late pre-emption. In order to avoid conflation with joint causation, one must add the requirement

that whether, how, or when E occurs does not depend on A or on C (even while it does depend on their sum). One must also require that there be no interruption of the causal process between A and E and likewise between C and E (note that this is the only difference between the case represented by [Fig. 8.5] and that represented by [Fig. 8.6]). Finally, one needs to add a requirement that A and C are distinct events, and that A does not depend on C and C does not depend on A. Moreover, all this must be done with non-circular definitions of 'interruption', 'causal process', and 'distinct'.[14] At this point, the prospects for a mereological sum treatment look grim, and it is unclear what sort of alternative treatments are available.

FURTHER READING

For early versions of counterfactual theories, see Lewis (1973*b*; 1986*c*), and Lyon (1967). For discussion of omissions, see Beebee (2004), McGrath (2005), Lewis (1986*c*), and Lewis (2004*b*). For discussion of problems with pre-emption, see Collins, Hall, and Paul (2004), Paul (1998*a*; 1998*b*), Lewis (1986*a*; 2004*b*), and Hitchcock (2007). For discussion of overdetermination and application to non-reductionism, see Paul (2007). For an in-depth cataloguing and assessment of these and other problems and issues involving counterfactual analyses of causation, see Hall and Paul (forthcoming).

REFERENCES

BEEBEE, H. (2004). 'Causing and Nothingness', in Collins, Hall, and Paul (2004).

BENNETT, K. (2003). 'Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It'. *Noûs*, 37/3: 471–97.

—— (forthcoming) 'Two Causal-isms'.

CHALMERS, D., and JACKSON, F. (2001). 'Conceptual Analysis and Reductive Explanation', *Philosophical Review* 100: 315–61.

COADY, D. (2004). 'Preempting Preemption', in Collins, Hall, and Paul (2004).

COLLINS, J., HALL, N., and PAUL, L. A. (2004). *Causation and Counterfactuals*. Cambridge, Mass.: MIT.

DOWE, P. (2000). *Physical Causation*. New York: Cambridge University Press.

ELGA, A. (2001). 'Statistical Mechanics and the Asymmetry of Counterfactual Dependence', *Philosophy of Science* 68/3 Suppl.: 313–24.

FAIR, D. (1979). 'Causation and the Flow of Energy', *Erkenntnis* 14: 219–50.

GANERI, J., NOORDHOF, P., and RAMACHANDRAN, M. (1996). 'Counterfactuals and Preemptive Causation', *Analysis* 56: 216–25.

GOOSENS, W. (1979). 'Causal Chains and Counterfactuals', *Journal of Philosophy* 76/9: 489–95.

HALL, N. (2000). 'Causation and the Price of Transitivity', *Journal of Philosophy* 97/4: 198–222.

—— (2004). 'Two Concepts of Causation', in Collins, Hall, and Paul (2004).

—— (2006). 'Philosophy of Causation: Blind Alleys Exposed; Promising Directions Highlighted', *Philosophy Compass* 1/1: 86–94.

—— and PAUL, L. A. (forthcoming) *Causation: A User's Guide*. Oxford: Oxford

University Press.

HALPERN, J., and PEARL, J. (2005). 'Causes and Explanations: A Structural-Model Approach. Part I: Causes', *British Journal for the Philosophy of Science* 56/4: 843–87.

HITCHCOCK, C. (2001). 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98: 273–99.

—— (2008). 'Prevention, Preemption, and the Principle of Sufficient Reason', *Philosophical Review* 116/4: 495–532.

—— (forthcoming). 'Conceptual Analysis Naturalized'.

—— and Knobe, J. (forthcoming). 'Norms and Causation'.

JACKSON, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.

KVART, I. (2001). 'The Counterfactual Analysis of Cause', *Synthese* 127/3: 389–427.

LEPORE, E., and LOEWER, B. (1987). 'Mind Matters', *Journal of Philosophy* 84/11: 630–42.

LEWIS, D. (1973*a*). *Counterfactuals*. Oxford: Blackwell.

—— (1973*b*). 'Causation', *Journal of Philosophy* 70: 556–67.

—— (1979). 'Counterfactual Dependence and Time's Arrow', *Noûs* 13: 455–76.

—— (1986*a*). *Philosophical Papers II*. Oxford: Oxford University Press.

—— (1986*b*). 'Events', in Lewis (1986*a*).

—— (1986*c*). 'Postscripts to "Causation"', in Lewis (1986a).

—— (2000). 'Causation as Influence', *Journal of Philosophy* 97/4: 182–97.

—— (2004*a*). 'Causation as Influence', extended version, in Collins, Hall, and Paul (2004).

—— (2004*b*). 'Void and Object', in Collins, Hall, and Paul (2004).

LOMBARD, L. B. (1986). *Events: A Metaphysical Study*. London: Routledge & Kegan Paul.

LYON, A. (1967). 'Causality', *British Journal for the Philosophy of Science*, 18: 1–20.

McDERMOTT, M. (1995). 'Redundant Causation', *British Journal for the Philosophy of Science* 46: 423–44.

McGRATH, S. (2005). 'Causation by Omission: A Dilemma', *Philosophical Studies* 123: 125–48.

—— (forthcoming). 'Uneventful Causation'.

MACKIE, J. L.(1965). 'Causes and Conditions', *American Philosophical Quarterly* 2/4: 245–64.

MELLOR, H. (1995). *The Facts of Causation*. London: Routledge.

—— (2004). 'For Facts as Causes and Effects', in Collins, Hall, and Paul (2004).

MENZIES, P. (2004). 'Difference-Making in Context', in Collins, Hall, and Paul (2004).

PAUL, L. A. (1998*a*). 'Problems with Late Preemption', *Analysis* 58/1: 48–53.

—— (1998*b*). 'Keeping Track of the Time: Emending the Counterfactual Analysis of Causation', *Analysis* 58/3: 191–8.

—— (2000). 'Aspect Causation', *Journal of Philosophy* 97/4: 297–328.

—— (2007). 'Constitutive Overdetermination', in J. Campbell, M. O'Rourke, and H. Silverstein (eds.), *Causation and Explanation*. Cambridge, Mass.: MIT, 265–90.

—— (forthcoming). 'The Handmaiden's Tale'.

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge

University Press.

SARTORIO, C. (2006). 'On Causing Something to Happen in a Certain Way without Causing It to Happen', *Philosophical Studies* 129/1: 119–36.

SCHAFFER, J. (2000*a*). 'Causation by Disconnection', *Philosophy of Science* 67/2: 285–300.

—— (2000*b*). 'Trumping Preemption', *Journal of Philosophy* 97/4: 165–81.

—— (2001). 'Causation, Influence and Effluence', *Analysis* 61/269: 11–19.

—— (2005). 'Contrastive Causation', *Philosophical Review* 114: 297–328.

SIDER, T. (2003). 'What's So Bad about Overdetermination?', *Philosophy and Phenomenological Research* 67/3: 719–26.

STALNAKER, R. (1968). 'A Theory of Conditionals', in N. Rescher (ed.), *Studies in Logical Theory*. Oxford: Blackwell.

STREVENS, M. (2003). 'Against Lewis's New Theory of Causation', *Pacific Philosophical Quarterly* 84/4: 398–412.

TOOLEY, M. (1987). *Causation: A Realist Approach*. Oxford: Oxford University Press.

—— (1990). 'Causation: Reductionism versus Realism', *Philosophy and Phenomenological Research* 50, Suppl.: 215–36.

WINSHIP, C., and MORGAN, S. L. (1999). 'The Estimation of Causal Effects From Observational Data', *Annual Review of Sociology* 25: 659–707.

WOODWARD, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

YABLO, S. (2004). 'Advertisement for a Sketch of an Outline of a Prototheory of Causation', in Collins, Hall, and Paul (2004).

CHAPTER 9
PROBABILISTIC THEORIES

JON WILLIAMSON

1. INTRODUCTION

Causal relationships are typically accompanied by probabilistic dependencies—normally when *A* causes *B* the former raises or lowers the probability of the latter. Probabilistic theories of causality usually try to characterize or analyse causality in terms of these probabilistic dependencies: they try to provide probabilistic criteria for deciding whether *A* causes *B*, and often maintain that causality just *is* the corresponding pattern of probabilistic relationships. This chapter provides an introduction to and criticism of such accounts. While it will be argued that probabilistic theories are ultimately unsuccessful, work on probabilistic causality has shed a great deal of light on the relationship between causality and probability and hence these theories repay a thorough understanding.

The chapter is organized as follows. In sect. 2 some key distinctions are introduced: these are helpful for categorizing theories of causality. The idea of a probabilistic theory of causality is discussed in sect. 3; then some of the key early theories are presented: Reichenbach's in sect. 4, Good's in sect. 5, and Suppes's in sect. 6. More recently, major steps have been taken by researchers in Artificial Intelligence and these have led to the causal net formalism (sect. 7), perhaps the most important contemporary probabilistic theory. The influence of this formalism is sketched in sect. 8. It is argued in sect. 9 that probabilistic theories flounder because they admit counterexamples, and because they fail to accommodate the important connection between causality and physical mechanisms (sect. 10). In sect. 11 the epistemic theory of causality is introduced. This theory, I claim, allows one to take advantage of the insights of probabilistic causality while avoiding its pitfalls.

2. CATEGORIZING PHILOSOPHICAL THEORIES OF CAUSALITY

There are a wealth of theories of causality and these theories can be categorized according to the way they answer a range of key questions.

Some questions concern the causes and effects that are related by causality. First, are the

causal relata single-case or generic? A philosophical theory of causality might hold that a cause or effect concerns only a single occasion and so either obtains or fails to obtain: for example, *Audrey's letter will cause Balthasar anguish when he reads it* . Or it might hold that causes and effects can obtain and fail to obtain on different occasions: *smoking causes cancer*. In the former case, causes and effects are called *single-case, particular*, or *token-level*; in the latter case they are *generic, repeatably instantiatable*, or *type-level*. Second, are the causal relata *population-level* or *individual-level*? In the former case a cause or effect concerns a group of individuals: *an increase in inequality of wealth in Britain in* 2007 *caused a reduction in happiness* (note that cause and effect are single-case here). On the other hand, an individual-level cause or effect concerns only one individual at a time: *viral infection causes an immune response* (cause and effect are generic in this example). Of course *all* these kinds of causal relata occur in our causal claims, apparently without any great problem, so any theory that considers one kind to the exclusion of the other kinds provides only a partial account of causality.

Other questions concern the causal relation itself. First, is causality some kind of *physical* connection between cause and effect? Or is it purely *mental* in the sense that it is a feature of some (possibly idealized) individual's epistemic state? Second, is the causal relation *objective*, in the sense that if two agents disagree as to causal relationships, at least one of them must be wrong, or is it *subjective*, admitting a degree of personal choice? Third, does the theory in question attempt to understand actual or potential causality? A golf ball bouncing off a tree is in general a preventative of it going into the hole, though it may, in fact, cause it to go into the hole. The general case is sometimes known as *potential causation* or *possible causation* while the factual case is called *actual causation*.

While these clearly do not exhaust the questions one might ask, they are often viewed as central questions in the debate about causality. It is certainly useful to be aware of the distinctions they draw when trying to understand an unfamiliar philosophical theory of causality.

3. PROBABILISTIC THEORIES OF CAUSALITY

Most probabilistic theories of causality are motivated by the following central intuitions: (1) changing a cause makes a difference to its effects, and (2) this difference-making shows up in probabilistic dependencies between cause and effect. (There are exceptions, e.g. the early Reichenbach as outlined in sect. 4.) Many proponents of probabilistic theories will go further by maintaining that probabilistic dependencies *characterize* the causal relation, i.e. provide necessary and sufficient conditions for causal connection, of the form: *C* causes *E if and only if* appropriate probabilistic dependencies obtain. They often go further still by maintaining that the probabilistic dependencies *analyse* the causal relation: '*C* causes *E*' *just means that* the corresponding probabilistic dependencies obtain.

When characterizing the causal relation the probabilistic dependencies may themselves be formulated using causal terms, but in an analysis of causality any reference to causal terms in the probabilistic conditions should in principle be eliminable. With a probabilistic characterization of causality, a characterization of probability must also be provided, and to

complete a probabilistic analysis of causality, one needs an analysis of probability. This gives rise to questions about probability: is probability physical or epistemic? Does probability arise from indeterminism? Indeed all the questions of sect. 2 have probabilistic analogues, and when assessing probabilistic theories of causality these probabilistic analogues need to be settled in order to answer the causal questions.

The key concern is to come up with a set of probabilistic conditions that yields a plausible probabilistic theory of causality. In the following sections we shall consider several important attempts to do just that. A number of specific problems have been raised in the literature for each of the attempts given; rather than wade through a slurry of technical concerns we shall focus, in sects. 9 and 10, on two very general problems that face probabilistic theories.

4. REICHENBACH

Russell (1913) offered a compelling critique of the notion of cause, arguing that the folk notion is incoherent in several respects and that the fundamental sciences do not appeal to causal terms but instead to functional equations. Following attacks on causal language by Mach (1905) and Pearson (1892, 3rd edn. 1911), this created a context of some wariness towards causality (cf. sect. 8).

Thus Reichenbach (1959, first published 1923) went against the grain by taking causal relationships seriously and interpreting them as the lawlike functional equations of physics. He argued against aprioristic and conventional views of causality, and in favour of his 'probabilistic conception of causality' (ibid. }}6–8). It should be emphasized, though, that Reichenbach's probabilistic conception of that book was not a metaphysical account of causality but instead an account of the epistemology of causality. Since most probabilistic accounts of causality are intended as metaphysical accounts, and since Reichenbach did at a later stage offer a metaphysical probabilistic account of causality, this is apt to cause some confusion.

The epistemological account of Reichenbach (1959) proceeded along the following lines. Causal discovery does *not* take the form: 'we notice that certain laws hold in *particular* instances, and we infer that the laws will hold in *all* instances' (ibid. 131). This is because we cannot suppose causal laws obtain in particular instances without other causal knowledge: 'if the probability that causality holds in a specific case could not be independently established for at least one case, it could not be increased by an appeal to other cases. Such an appeal presupposes the probability that causality holds has been determined for these other cases' (ibid. 132). *No causes in, no causes out,* to use a contemporary slogan. Instead, 'the correct inference has the following form: since we have observed that the same function governs a finite number of observations, we conclude that it governs *all* observations' (ibid.). Reichenbach's account is probabilistic in the sense that it attaches probabilities to functional equations and thus to causal laws, not in the sense that the causal laws themselves are reducible to probability relationships.

The approach of the later Reichenbach (1971, first published 1956) is very different in that it attempts to analyse causality itself: 'in the present book, I wish to study the cause–effect relation in itself; that is, to treat it no longer as primitive, but to reduce it to other relations'

(ibid. 25). Causality is no longer viewed just as a functional relationship, because functional equations are symmetric in cause and effect while causal relations are asymmetric (ibid. 28). Reichenbach (ibid. sect. 19) attempts to analyse the direction of time in terms of the direction of causality, and appeals the following Principle of the Common Cause to determine the direction of causality:

If coincidences of two events A and B occur more frequently than would correspond to their independent occurrence, that is, if the events satisfy relation $[P(AB) > P(A)P(B)]$, then there exists a common cause C for these events such that the fork ACB is conjunctive, that is, satisfies relations $[A \perp\!\!\!\perp B|C, P(A|C) > P(A|\bar{C}), P(B|C) > P(B|\bar{C})]$. (ibid. 163)

In this quotation, as elsewhere in this chapter, I have substituted contemporary mathematical notation for the original notation: $A \perp\!\!\!\perp B \mid C$ means that A and B are probabilistically independent, conditional on C. The latter relations imposed by Reichenbach imply that $P(AB) > P(A)P(B)$, so the common cause can be said to explain the dependence between A and B.

Reichenbach's key idea was that the Principle of the Common Cause should be used to determine the direction of causality and thus the direction of time: where there is a fork ACB (see Fig. 9.1) such that the relations in the Principle of the Common Cause hold, and where there is no other C' that satisfies these conditions with respect to A and B, then C is the common cause of A and B and is earlier than A and B.

Reichenbach (1971: §22) extends the analysis to deal with causal betweenness: C is *causally between A and B* (as in Fig. 9.2) if $1 > P(B \mid C) > P(B|A) > P(B) > 0, 1 > P(A|C) > P(A|B) > P(A) > 0$ and $A \perp\!\!\!\perp B \mid C$.

The aim is that these probabilistic conditions can be used to determine a complete causal graph. Here 'determine' is used in a metaphysical rather than epistemological sense: the conditions are not a description of how one ought to discover causal relationships but rather they offer a characterization and analysis of the causal relation. According to this characterization, a cause raises the probabilities of its direct effects, and no other event renders the cause and a direct effect probabilistically independent.

Reichenbach had developed his own frequency theory of probability and these frequencies were supposedly non-causal, so causality is ultimately analysed in terms of frequency relations. [1] In terms of the distinctions of sect. 2, causality is thus a physical and objective relation; since frequencies handle generic attributes, Reich-enbach's analysis deals only with generic relata and potential causation (leaving open the question as to whether the causal relata are population-level or individual-level).
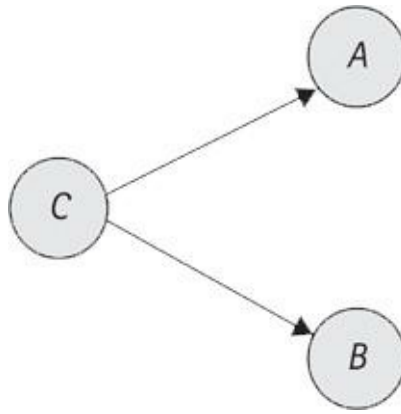
Fig. 9.1 A fork or v-structure.
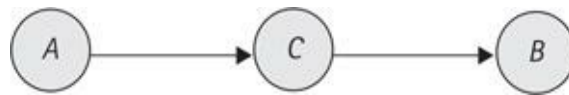
Fig. 9.2 *C* is causally between *A* and *B*.

Note that Reichenbach represented causal connections graphically, using diagrams like those depicted in Figs. 9.1 and 9.2. He called these diagrams *causal nets*. Figure 9.2, which represents a sequence of events with a direct causal connection from one event to the next, is called a *causal chain*. This graphical representation has become an important feature of recent work on probabilistic causality, as will be explained in sect. 7.

Reichenbach's probabilistic analysis of causality was criticized by Salmon (1980: sect. 2), although in later works Salmon (1984; 1998) went on to develop the mechanistic ideas to be found in Reichenbach's writings. See Martel (2000) for a defence of Reichenbach's approach.

5. GOOD

Good (1959: 307) attempted to provide an analysis of causality in terms of Popperian propensity, a kind of physical probability.[2] As with Reichenbach, Good avoids reference to time in order to leave open the possibility of analysing time in terms of causality. For an event F to be a cause of event E, Good (1959) requires that (a) E and F both occur; (b) $P(E|H)1$ and $P(F|H)1$ where H consists of all the laws of nature (H_1) and all the true background conditions that are taken for granted (H_2); (c) $P(E|FH) > P(E|\bar{F}H)$; (d) FH_2 does not logically imply E; (e) there is an event G that did not occur, that could have prevented F, $P(\bar{F}|GH) \simeq 1$, and whose absence would not ensure F, $P(F | \bar{G}H)1$; and (f) $P(E|GH) \simeq P(E|\bar{F}GH) \simeq P(E| \bar{G}FH) \simeq P(E| \bar{F}H)$. Good ultimately rejected this account on the grounds that conditions (e) and (f) in fact hold trivially in the presence of the other conditions (a–d). This led Good to seek a better probabilistic account.

Good (1961*a*; 1961*b*) provided an updated account. Here Good no longer avoids appealing to time in the analysis of causality, explicitly requiring that a cause be earlier than its effects. This account analyses causality in terms of physical probability, but it goes beyond his previous account in that it attempts to give quantitative measures of both potential and actual causation. The (potential causal) *tendency of F to cause E* is measured by

$$\log \frac{P(\bar{E}|\bar{F}H)}{P(\bar{E}|FH)}$$

where *H* consists of all laws of nature and the background conditions before *F* started. Thus for *F* to be a potential cause of *E*, the two must be probabilistically dependent conditional on *H*. The (actual causal) *degree to which F caused E* is the limit, as the sizes of the events tend uniformly to zero, of the strength of the network of causal connections between *E* and *F*. Here the strength of a link from *F* to *E* is measured by the tendency of *F* to cause *E*; the strength of the network as a whole is a function of these link strengths which takes into account interactions amongst causes.

Good quite explicitly develops the notion of a *causal net*, which is represented by a directed acyclic graph whose nodes are events and whose arrows signify causal connections from cause to effect and chart the probabilistic dependencies that obtain amongst the events. Good's approach is devised to be able to cope with token events, and clearly construes causality as an objective, physical relation. Good preferred his own account over Reichenbach's on the grounds that Reichenbach's account turns out to be vacuous: for event *F* to be a cause of event *E*, Reichenbach requires that *F* raise the probability of *E* and that there is no event *G* that renders *E* and *F* probabilistically independent; Good (1961*b*: app. 1) argued that one can always gerrymander some event *G* that renders the two variables probabilistically independent, in which case there are no causal relations at all under Reichenbach's account.

Good's account of causality is criticized in Salmon (1980: }1 and 1988).

6. SUPPES

Unlike the later Reichenbach but like the later Good, Suppes (1970) appeals to time in his analysis of the causal relation. Suppes (1970: 12) has three central steps to his analysis. First he introduces the notation A_t to signify that event *A* occurs at time *t*, and gives a preliminary definition: 'the event $B_{t'}$ is a prima facie cause of the event A_t if and only if (i) $t' t$, (ii) $P(B_{t'}) > 0$, (iii) $P(A_t|B_{t'}) > P(A_t)$'. Suppes (ibid. 25) then defines a prima facie cause $B_{t'}$ of A_t to be a *spurious cause* of A_t if there is a prior partition $\pi_{t''}$ of events that screens off *B'* from A_t: for all elements $C_{t''}$ of $\pi_{t''}$, (i) $P(B_{t'} C_{t''}) > 0$, (ii) $P(A_t|B_{t'} C_{t''}) = P(A_t|C_{t''})$. Finally, a *genuine cause* is a prima facie cause that is not spurious.

Suppes goes on to define *direct cause* and *negative cause*, and then extends his account to deal with causal relations between variables, not just events. Suppes is clear that he intends his account as an analysis of causality, rather than merely a characterization of the causal relation.

He leaves open the interpretation of probability (ibid. }2, pp. 79–80) and intends that the account apply to both generic and single-case relata (ibid. 79). Suppes does not think that causality is entirely objective, in the sense that the causal relation is not uniquely determined. First, causal relationships are relative to the conceptual framework (the set of events or variables) under consideration (ibid. 75). Second, 'the analysis of causes is always relative to a particular conception of mechanism, and it does not seem satisfactory to hold that the analysis of mechanism is ever complete or absolute in character' (ibid. 72). For example, the analysis depends on whether time is continuous or discrete. While Suppes discusses mechanisms in some detail, he is not explicit about the way mechanisms fit into the analysis. The connection with mechanisms would suggest that Suppes considers the causal relation to be a physical relationship; however, the fact that he admits the possibility of a mental interpretation of probability suggests that a mental notion of cause may also be adopted. In sum, while Suppes offers a single formal reduction of causality to probabilities, he is a pluralist about causality— causality varies according to the interpretation of probability, the conceptual framework, and the notion of mechanism under consideration.

See Salmon (1980: §3) for critical discussion of Suppes's probabilistic account of causality.

Note that Salmon (1984: 190) explicitly advocates 'probabilistic causality', but under Salmon's view causality is probabilistic only in the sense that it is not deterministic (*not* in the sense that causal relations can be characterized or analysed in terms of probabilistic relations):

I cannot think of any reason to suppose that ordinary causal talk would dissolve into nonsense if Laplacian determinism turned out to be false. I shall therefore proceed on the supposition that probabilistic causality is a coherent and important philosophical concept.

In advocating the notion of probabilistic causality, neither Suppes nor I intend to deny that there are sufficient causes; indeed, Suppes explicitly introduces that concept into his theory (1970, p. 34, def. 9). On our view, sufficient causes constitute a limiting case of probabilistic causes. On the sufficiency/necessity view, which we reject, this limiting case includes all bona fide cause–effect relations. The latter approach to causality seems needlessly restrictive.

Thus Salmon's 'probabilistic causality' is a reaction to Mackie's account of causality in terms of *inus* conditions (which holds that a cause is an Insufficient but Necessary component of a condition that is Unnecessary but Sufficient for an effect—see Ch. 7 on regularity theories of causation in this volume).

Suppes's account was developed by Cartwright (1979) who put forward the principle:

CC: *C* causes *E* iff $P(E|CK) > P(E|K)$ for all states *K* of the *E*'s other causes that are not between *C* and *E*.

Cartwright did not take this to be an analysis of causality because cause appears on both sides of 'iff'. The *context unanimity* of this condition was relaxed by Skyrms (1980: 108–9) who claimed that a cause need only raise the probability of the effect relative to *some* state *K*, rather than all such states (though there should be no state *K* for which *C* lowers the probability of *E*). Eells and Sober (1983) discussed Cartwright's condition and argued that this

kind of condition only holds of generic causal claims, not of single-case claims. Eells (1991) provided a detailed defence of generic causality using this type of probabilistic account, and went on to provide a probabilistic theory of single-case causality, appealing to *probability trajectories*, which trace the evolution of single-case probabilities over time.

7. CAUSAL NETS

The 1980s saw a very influential mathematization of the notion of cause. This line of work stemmed from efforts among Artificial Intelligence researchers to automate reasoning in the face of uncertainty. In the 1970s so-called *expert systems* were developed in order to automate reasoning tasks that would normally require human expertise. For example, expert systems were developed for the diagnosis of intestinal problems and for mineral prospecting. Early expert systems tended to take the form of *rule-based systems*: the expert knowledge was encoded by logical rules and on inputting a set of facts (e.g. a patient's symptoms) the rules would be applied to generate inferences (e.g. diagnoses). It was soon realized that rules for diagnosis and other common expert tasks are rarely exceptionless, and that expert systems needed a way of handling this uncertainty. One way is to invoke the notion of probability: a particular pathology is more or less probable given a patient's symptoms, though rarely certain or certainly ruled out. In lieu of tractable computational procedures for handling probabilities, several non-probabilistic formalisms for handling uncertainty were developed. Then, in the 1980s, a formalism was developed which at one fell swoop offered the possibility of tractably representing and reasoning with probabilities on the one hand, and representing and reasoning with causal connections on the other. This is the formalism of Bayesian networks (Pearl 1988; Neapolitan 1990).

A Bayesian network consists of a directed acyclic graph whose nodes are variables in the domain of interest, together with the probability distribution of each variable conditional on its parents in the graph (or its unconditional distribution if the variable has no parents). The graph and the probabilities are tied together by a fundamental assumption known as the *Markov condition*: each variable is probabilistically independent of its non-descendants conditional on its parents in the graph, written $A_i \perp\!\!\!\perp ND_i \mid Par_i$ for each i. Then one can calculate any probability involving the variables in the domain via the identity $p(a_1 \ldots a_n) = \prod_{i=1}^{n} P(a_i|par_i)$ $P(a_i|par_i)$ where each a_i is an assignment of a value to variable A_i and par_i is the assignment of values to its parents Par_i consistent with $a_1, \ldots, a_n$.

For example, suppose there are four two-valued variables, *A, B, C, D*. A Bayesian net can be formed by taking the directed acyclic graph of and specifying the probability distribution of each variable conditional on its parents, e.g.

$$P(a^1) = .3, \; P(a^0) = .7$$
$$P(b^1) = .9, \; P(b^0) = .1$$
$$P(c^1|a^1b^1) = .4, \; P(c^0|a^1b^1) = .6$$
$$P(c^1|a^0b^1) = .5, \; P(c^0|a^0b^1) = .5$$
$$P(c^1|a^1b^0) = .1, \; P(c^0|a^1b^0) = .9$$
$$P(c^1|a^0b^0) = .2, \; P(c^0|a^0b^0) = .8$$
$$P(d^1|c^1) = .8, \; P(d^0|c^1) = .2$$
$$P(d^1|c^0) = .7, \; P(d^0|c^0) = .3$$

Then for instance

$$
\begin{aligned}
P(a^1b^0c^1d^1) &= P(a^1)P(b^0)P(c^1|a^1b^0)\,P(d^1|c^1) \\
&= .3 \times .1 \times .1 \times .8 = .0024.
\end{aligned}
$$

In recent years a whole host of techniques has been developed for efficiently calculating probabilities from a Bayesian net and for constructing a Bayesian net to match the probability distribution of a dataset.
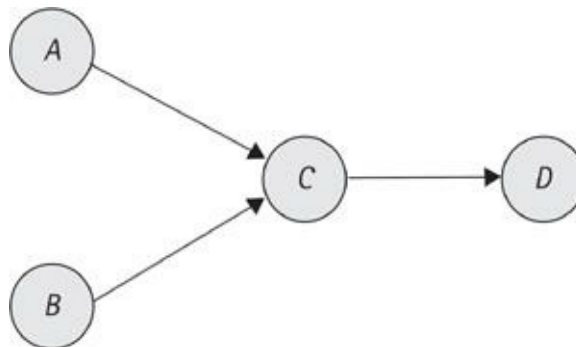
Fig. 9.3 An example of a directed acyclic graph.

Bayesian nets are often used to represent and reason with causal relationships. A *causally interpreted Bayesian net* or *causal net* is a Bayesian net in which the arrows of the graph are interpreted as denoting direct causal relationships. Thus for example under a causal interpretation Fig. 9.3 says that A and B cause C and also cause D indirectly via C. Under a causal interpretation, the Markov Condition—now called the *Causal Markov Condition*—says that each variable is probabilistically independent of its non-effects conditional on its direct causes. It is normally assumed that if the graph in the net correctly portrays the causal relationships amongst the variables, and no causally relevant variable is omitted, then the Causal Markov Condition must hold, though in Williamson (2005) I argue that this condition can at best be justified as a *default* assumption.

Causal nets can be invoked as a characterization of the causal relation as follows. Suppose physical reality is conceptualized using some domain V of variables, and that P^* is the physical probability function (propensity or frequency function) over this domain. Let C^* be

the smallest directed acyclic graph on *V* that satisfies the Markov Condition with respect to *P**. Then, it may be claimed, C^* characterizes the causal relationships. This type of characterization is invoked by many advocates of causal modelling, e.g. Spirtes, Glymour, and Scheines (1993) and Pearl (2000), and is very close to those developments of Suppes's account which do not invoke context unanimity (sect. 6). Often restrictions are made on *V* (e.g. that *V* should pick out all and only the physical events) and further assumptions are made concerning *P** (e.g. that all the independencies of *P** are representable by some directed acyclic graph; that *P** is defined over interventions as well as observations) in order to ensure that C^* is independent of *V* and is uniquely determined. The details of this formalism can be found in Ch. 14 on causal modelling in this volume.

Some proponents of this kind of characterization go further by using it to analyse the causal relation itself (see e.g. Spohn 2002). Under this view, the causal relation *just is* the relation picked out by C^*: causal relationships are a chart of the independencies satisfied by physical probability.

8. CAUSAL TALK

Probabilistic causality—in particular the causal net formalism—provides a mathematical calculus for handling causality that has been enormously influential. Prior to the late twentieth century 'cause' was a dirty word to most research scientists, who, since the time of Pearson and Russell, had been warned that correlation need not imply causation and that causality is a muddled notion devoid of mathematical treatment. But the increasing mathematization of causality that accompanied the rise of probabilistic causality began to change all this. The causal net formalism reached a wide audience with the books of Pearl (1988) and Neapolitan (1990) and this dissemination helped bring causal talk back in vogue.

Evidence of this transition can be obtained by analysing titles of books and research papers. Figure 9.4 shows a decline in causal talk in the nineteenth century and into the twentieth century: the bars show the proportion of published books in the English language with a word in the title beginning with 'caus-' (from the British Library database). The proportion is of course very small, so percentage figures are omitted, but the general downward trend can be observed. It can be seen, though, that this trend lasts only until the 1970s. Figure 9.5 shows the period since the 1970s in detail: here the bars show the proportion of published research papers with a word in the title beginning with 'caus-' (from the Web of Science database). This graph shows a steady increase in causal talk.

While this rise might be attributable to the influence of probabilistic causality and causal nets, another possible explanation of this phenomenon is that the rise is due to an increase in the number of papers in medicine, a subject that is inherently causal.[3] This is a plausible hypothesis because, as Bauer (1998) shows, from the 1930s to the 1990s there were more articles about physics than medicine in quality newspapers, but in the 1990s medicine overtook physics. Perhaps, then, more research is being done in medicine, which uses more causal talk. But Fig. 9.6 shows that while the portrayal of science by the media may have changed around 1990, the balance of science did not itself change. Here the black bars show the numbers of papers in journals whose titles contain a word beginning with 'med-' while the

shaded bars show the number of papers in journals whose titles contain a word beginning with 'phys-'. While the numbers of papers are growing year by year, it is clear that research in medicine is not overtaking research in the physical sciences. shows the same phenomenon, but with 'bio-' journals compared to 'phys-' journals.
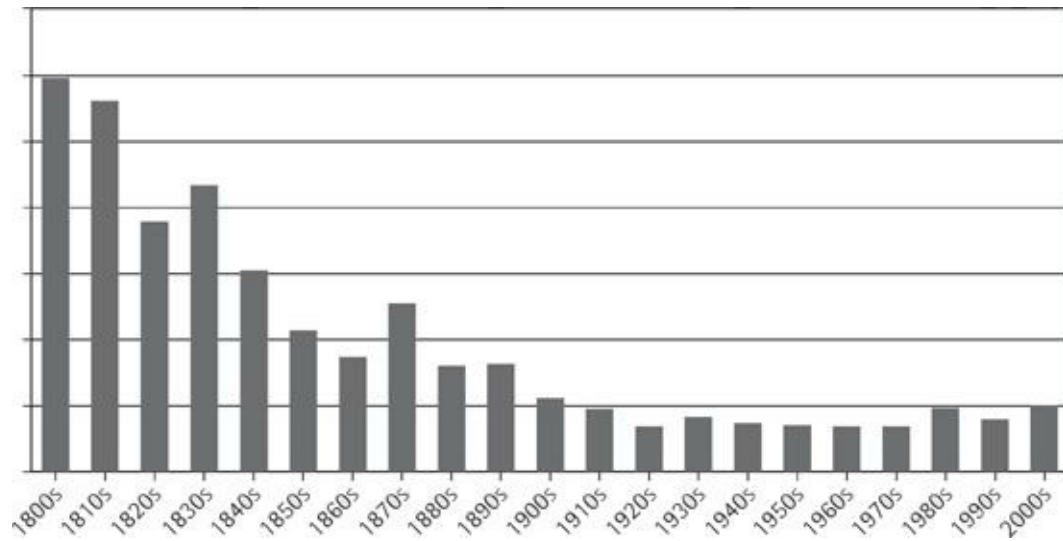
Fig. 9.4 Proportion of books whose title includes causal terms.

In sum, causal talk is coming back into vogue and this is no doubt at least partly attributable to the rise of probabilistic causality and causal net modelling.

9. COUNTEREXAMPLES TO PROBABILISTIC CAUSALITY

In this section and the next we shall take a look at two general kinds of worry about the probabilistic approaches to causality outlined above. In this section we shall see that the connections between probability and causality that are invoked by the above theories admit counterexamples. In sect. 10 it is argued that physical mechanisms are neglected by the above accounts but are crucial to our understanding of causality.
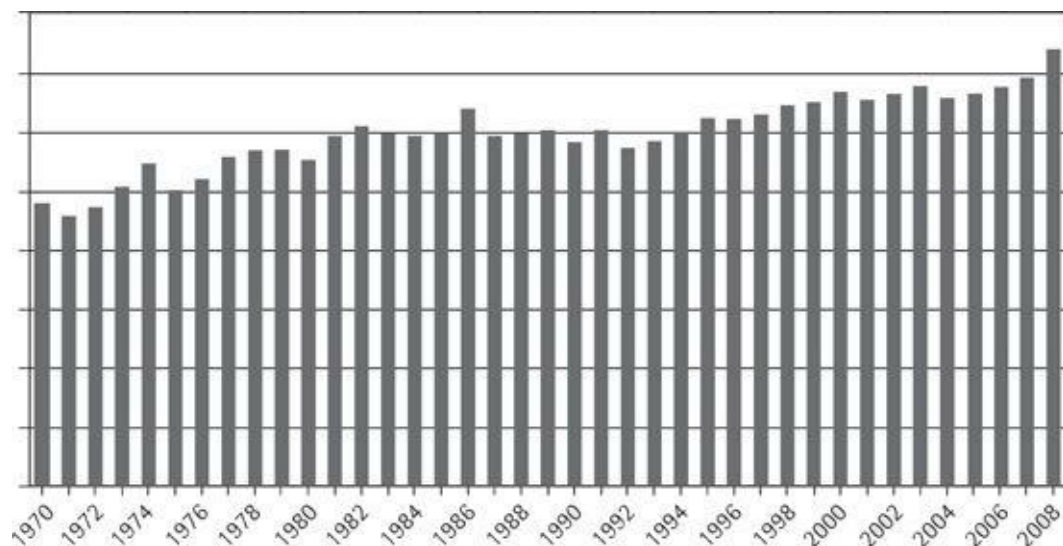
Fig. 9.5 Proportion of papers whose title includes causal terms.
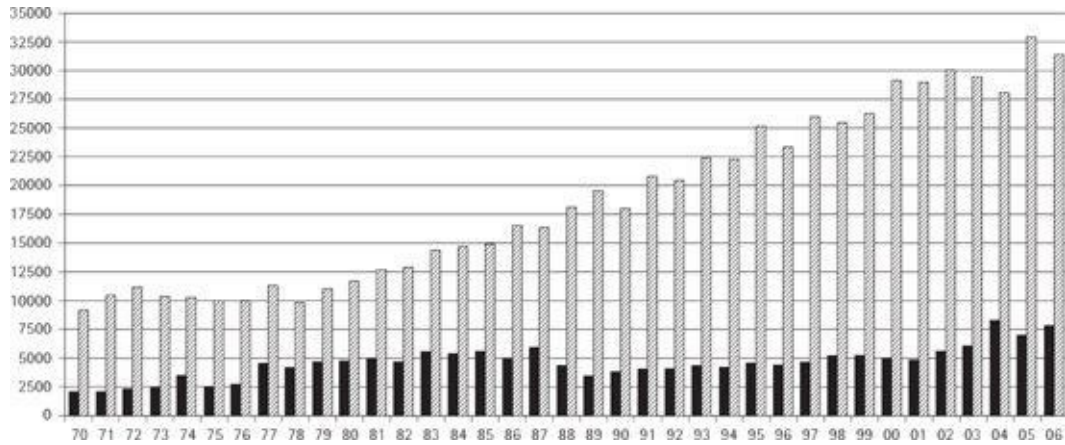
Fig. 9.6 Numbers of medical (black) and physical (shaded) papers.
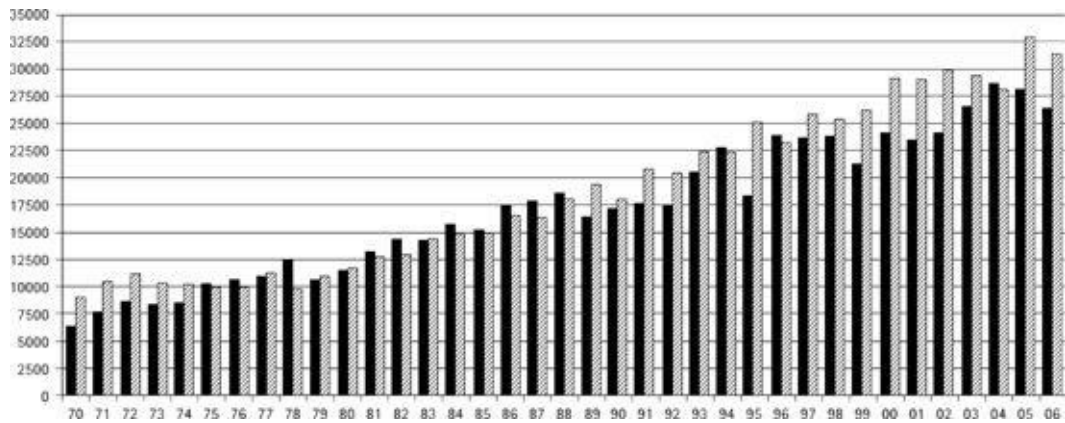
Fig. 9.7 Numbers of biological (black) and physical (shaded) papers.

Note that the Causal Markov Condition implies the following version of the Principle of the Common Cause:

PCC: If variables A and B are probabilistically dependent then one causes the other or there is a set U of common causes in V which screen off A and B, i.e. render them probabilistically independent, $A\,B\,j\,U$.

This version of the Principle of the Common Cause is also a consequence of Reichenbach's own version (under a suitable mapping between events and variables) and of developments of Suppes's account that do not invoke context unanimity.[4]

Unfortunately, PCC is oversimplistic. It says that any probabilistic dependence can be fully accounted for by causal connections. In fact, though, probabilistic dependencies may be

attributable to other kinds of relationships between the variables. A and B may be dependent not because they are causally related but because they are related logically (e.g. where an assignment to A is logically complex and logically implies an assignment to B), mathematically (e.g. mean and variance variables for the same quantity are connected by a mathematical equation), or semantically (e.g. A and B are synonymous or overlap in meaning), or are related by non-causal physical laws or by domain constraints. In such cases there may be no common cause to accompany the dependence, or if there is, the common cause may fail fully to screen off A from B. To take a simple example, if a logically implies b then $P(b|a) ¼ 1$ while $P(b)$ may well be less than 1. In such a case variables A and B (where A takes assignments a and $Ø a$ and B takes assignments b and $Ø b$) are probabilistically dependent; however it is rarely plausible to say that A causes B or vice versa, or that they have a common cause. Further examples can be found in e.g. Williamson (2005: §4.2). A typical response to this objection takes the form: as long as you take care to ensure that your variables are not related logically, mathematically, semantically etc. then PCC will hold. But this is to render PCC devoid of content, for it is tantamount to saying: as long as you ensure that there are no probabilistic dependencies that are not attributable to causal relationships, then all dependencies will be attributable to causal relationships. Moreover, this response renders PCC much less useful as an epistemological tool, for it is often very hard to tell whether a dependency is attributable to a non-causal relationship.

PCC also hinges on the particular physical interpretation of probability that is adopted. Under an *actual frequency* interpretation of probability, probabilistic dependencies may be entirely accidental, having no underlying explanation and in particular no causal explanation. For example, consider observations of vehicles at a particular road junction today; it may be the case that the proportion of green vehicles that turn right happens to be larger than the proportion of red vehicles that turn right. These proportions are probabilities and hence the direction of turn and the colour of vehicle are probabilistically dependent. This dependency is not attributable to a common cause—it is accidental and in no need of explanation at all. In order to save PCC one might switch interpretation of probability, perhaps maintaining that if one were to have examined vehicles indefinitely today then direction of turn and colour would be independent with respect to *limiting relative frequency*. But—aside from reservations one might have about the move to a counterfactual account of physical probability—it is a purely metaphysical assumption that accidental dependencies will disappear in the limit, and by no means a plausible assumption if one considers time series (Yule 1926; Sober 1988, 2001; Reiss 2007). As Sober observes, British bread prices and the Venetian sea level are correlated, not because they are causally connected but simply because they are both increasing over time for quite separate reasons.

10. MECHANISMS AND PROBABILISTIC CAUSALITY

In some cases causal relationships are not accompanied by the raising of probabilities. Consider the following example, adapted from examples of Salmon (1984: 196–202) and Dowe (2000: §II.6) and presented in more detail in Williamson (2005: §7.3). A potentially unstable atom can decay as follows: radioactive isotope a[1] will decay to either b^1 or b^0 each of which will decay to either c^1 or c^0; on the other hand isotope a^0 is stable and will not decay.

Moreover c^1 and c^0 can only be obtained by decay from b^1 or b^0 and b^1 and b^0 can only be obtained from a^1. We have that $P(c^1|b^1) = ¾$ (so $P(c^0|b^1) = ¼$), $P(c^1|b^0) = ¼$, $P(b^1|a^1) = ¼$, $P(b^1|a^0)$ ¼ 0, $P(a^1) = ½$. Let variable A take assignments a^1 or a^0, B take assignments b^1 or b^0, and C take c^1 or c^0. Then the causal picture is depicted in . Clearly b^0 is a potential cause of c^1 (and may be the actual cause of c^1) even though it lowers the probability of c^1.

Causal relationships need not even be accompanied by probabilistic dependencies. Consider the same example, but where $P(c^1|b^1) = ½$ and $P(c1|b^0) = ½$. In this case B is still a cause of C—indeed the only cause of C—but C and B are probabilistically independent. This is a serious problem for the probabilistic accounts of causality considered above, all of which require that causal relationships be coextensive with certain probabilistic dependencies. If causal relationships are not characterizable in terms of probabilistic relationships, that clearly puts paid to a probabilistic analysis of causality.

In these examples it is not the probabilities that tell us about the causal relationships but rather physical knowledge—knowledge about the physical entities and physical mechanisms that link them. One might think, then, that probabilistic accounts of causality should be abandoned in favour of mechanistic accounts—indeed, this was the line of argument taken by Wesley Salmon (see also Ch. 10 below on mechanistic accounts ('Causal Process Theories')).

But this is too quick. Mechanistic accounts fare no better, for there are cases in which it is clearly not mechanisms that tell us about causal relationships, but rather probabilistic dependencies. One important problem for mechanistic theories is to do with *absences*. There can hardly be mechanisms linking non-existent entities, yet absences are prime examples of causes and effects: e.g. missing the ferry causes Celia to be absent from work. Thus causal relationships can occur without mechanisms (though with accompanying probabilistic dependencies). Another important problem for mechanistic theories occurs where there is a mechanism of the appropriate sort but no causal relationship. There are any number of mechanisms linking a beating heart with liver failure, in cases where such failures occur. Many of these mechanisms qualify as causal mechanisms according to contemporary mechanistic theories of causality, yet one would not want to say that a beating heart is a cause of liver failure, because there is no probabilistic dependence between the two. Instead one would cite those events that are mechanistically connected to liver failure *and* which make a difference to the probability of liver failure as the causes of liver failure. A third problem for mechanistic theories is that mechanisms are heterogeneous across the sciences—mechanisms in physics are quite different from mechanisms in economics, for example; a physical-mechanism account is unlikely to cope with causal relations in economics, while an economic-mechanism account is unlikely to cope with causes in physics, and an account which cashes out causes in physics in terms of physical mechanisms and causes in economics in terms of economic mechanisms is unlikely to be able to explain the apparent homogeneity of the causal relation.

In sum, an account that seeks to characterize or analyse causality just in terms of probabilistic dependencies, or just in terms of physical mechanisms, will be inadequate. This conclusion led Hall (2004) and others to argue in favour of a pluralistic account of causality—sometimes causal relationships are probabilistic, in other cases they are mechanistic. But causal pluralism, discussed later in this volume, has its own set of problems (Russo and

Williamson 2007; Williamson 2006*a*). Pluralism has trouble accounting for the uniformity of our causal talk—there is no apparent ambiguity when I say 'smoking causes cancer', and it would not make sense to require that I clarify my claim by saying 'smoking mechanistically causes cancer' or 'smoking probabilistically causes cancer' since both mechanisms and probabilities are important evidence for this causal claim. This latter point is worth spelling out. If 'smoking causes cancer' is to be understood in terms of a probabilistic relationship between smoking and cancer, then there is an epistemological problem: it is hard to explain why, given that there was excellent probabilistic evidence in favour of smoking being a cause of cancer, the causal claim was not generally accepted until a plausible physiological mechanism linking smoking and cancer was discovered (Fisher, for one, argued that a physiological mechanism had to be found before the causal claim could be substantiated). If the causal claim is probabilistic, why should evidence of mechanisms be required over and above evidence of probabilities? On the other hand if the causal claim is mechanistic, it is hard to explain why, were a plausible mechanism known, one would normally still require evidence that the cause made a difference to the effect before the claim could be said to be substantiated. This epistemological problem besets probabilistic, mechanistic, and pluralist theories of causality alike—under a pluralist account 'smoking causes cancer' must be given one or other interpretation, but then it is not clear why evidence both of mechanisms and of probabilistic dependence is required.
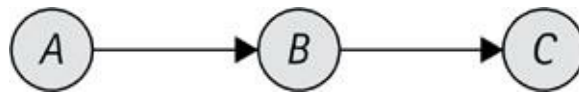
Fig. 9.8 Radioactive decay example.

One reason why mechanistic evidence is often required over and above evidence of probabilistic dependence is that causal claims need to be explanatory. Causal claims have two kinds of use: an *inferential* use, for making predictions, diagnoses, and strategic decisions, and an *explanatory* use, to give an account of why an effect occurred. In order to be put to the inferential use, it is crucial that a cause make a difference to its effect—that is, that cause and effect are probabilistically dependent—otherwise one could not predict the effect and could not intervene on the cause to produce the effect. But difference-making is not enough for the explanatory use of causal claims. It is not enough when asked 'why did the effect occur?' to answer 'because an event that makes a difference to the effect occurred', because that is no explanation at all—it still leaves the question 'why did the event make a difference to the effect?' In order to explain an event one needs to invoke some kind of theoretical knowledge —facts about the events and their linking mechanisms. Hence causal claims need to be associated, where possible, both with probabilistic dependencies and with mechanisms.[5]

Reichenbach thought that his probabilistic notion of causality is coextensive with a mechanistic account (Reichenbach 1971; Otte 1986) but the above examples show that this is not the case. It is hard to deny that both probabilities and mechanisms are important to our understanding of causality, yet it seems that probabilistic, mechanistic, and pluralist accounts all fail. There remains the thorny question of how to provide a viable account of causality that integrates probabilistic and mechanistic considerations.

11. THE EPISTEMIC THEORY OF CAUSALITY

I will argue that this question can be answered if we interpret causality as an *epistemic relation*. According to this view, causality is to be analysed neither in terms of physical probabilities nor in terms of physical mechanisms, but in terms of an agent's epistemic state. This type of view can be traced to Hume and Kant (Williamson 2005: §9.2) and also Mach:

There is no cause nor effect in nature; nature has but an individual existence; nature simply *is*. Recurrence of cases in which *A* is always connected with *B*, that is, like results under like circumstances, that is again, the essence of the connection of cause and effect, exist but in the abstraction which we perform for the purpose of mentally reproducing the facts. (Mach 1883: 483).

Much of the authority of the ideas of cause and effect is due to the fact that they are developed *instinctively* and involuntarily, and that we are distinctly sensible of having personally contributed nothing to their formation. We may, indeed, say, that our sense of causality is not acquired by the individual, but has been perfected in the development of the race. Cause and effect, therefore, are things of thought, having an economical office. (ibid. 485)

Thus to say that the causal relation is an epistemic relation is to say that causality is a feature of the way we represent the world rather than a feature of the agent-independent world itself.

An epistemic theory of causality can be developed using the following recipe.

First, take an ideal causal epistemology. In particular consider the way that one's evidence (including one's background knowledge) should constrain the causal beliefs that one has. (Here causal beliefs are not to be construed as beliefs *about* causality. Instead a causal belief is a certain *type* of belief, namely a directed relational belief, representable using directed acyclic graphs, and one that is put to the inferential and explanatory uses that typify causal reasoning.) Certain causal belief graphs are compatible with evidence, others are ruled out. The uses to which causal beliefs are put determines this mapping from evidence to a set of possible causal belief graphs. It is this ideal mapping that constitutes the required ideal causal epistemology.

I won't say much here about this mapping. In practice the epistemology of causality is much less controversial than the metaphysics of causality, and I shall just presume that this mapping is well defined. Current methods for causal discovery offer approximations to this ideal causal epistemology—one such can be found in Williamson (2006*a*: app. A) and there are many others. Presumably as science progresses the approximations to the ideal mapping will improve. What is clear from sects. 9 and 10, though, is that evidence both of probabilistic dependencies and of mechanisms will play a role in constraining the set of viable causal belief graphs. (The question of exactly how these are to be understood—which notion of mechanism

and which physical interpretation of probability is required—may be left to the ideal causal epistemology to decide.)

The second step of the recipe for epistemic causality is to take all empirical facts. By this I mean all facts about physical reality, but not facts about rational epistemology—for example, neither facts about the ideal causal epistemology, nor facts about causality itself (we shall turn to the question of what these facts are shortly).[6]

Third, apply the causal epistemology to this ideal set of evidence. This results in a set of ideal causal belief graphs. If a rational agent had as evidence the ideal evidence set, knew the ideal causal epistemology, and were able to apply the latter to the former then her causal beliefs would be representable by one of these ideal causal belief graphs. This set of ideal causal belief graphs characterizes the causal relation. If this set is a singleton, the causal relation is fully objective, otherwise there is some subjective choice as to what the causal relationships are—the extent of this subjective choice is proportional to the cardinality of the set of ideal causal belief graphs. The facts about causality are just the facts about *all* the graphs in the ideal set. Thus it is fact that A causes B just if A causes B in each ideal causal belief graph.

Fourth, extend this characterization to an analysis of causality. Causal relationships just *are* the result of applying the ideal causal epistemology to the ideal evidence set. They are the set of causal beliefs one should have were one to know all physical facts and the ideal causal epistemology and were one able to apply the latter to the former. (Here, as before, it must be emphasized that a causal belief is a kind of belief, a relational belief that is put to the inferential and explanatory uses that are associated with causal claims. It should not be confused with a belief about causality, which is a non-relational belief about the set of ideal causal graphs.) It is thus the uses to which causal claims are put that determines the nature of causality itself.

All this is obviously highly idealized, and while it may satisfy our need for a metaphysical account of causality, the application of an ideal causal epistemology to an ideal evidence set doesn't say much about how we can discover causal relationships in practice. The best we can do, of course, is to apply the causal epistemology of the moment to the evidence of the moment. Quite plausibly, some of our more entrenched causal claims of the moment will remain entrenched as our evidence and causal epistemology improves. Thus there is no reason why much of what we think of now as knowledge of causal relationships should not in fact be such knowledge (i.e. knowledge of the ideal causal belief graphs).

Note that steps 1–3 of this recipe are rather uncontroversial. Proponents of a probabilistic or mechanistic analysis of causality will surely agree with the claim that the application of the ideal causal epistemology to the ideal evidence set will characterize the causal relation— though of course they may differ as to what might constitute the ideal causal epistemology. Indeed proponents of any view of causality in which causal relationships are not radically unknowable will concur. Substantial disagreement only comes at step 4—the analysis. The proponent of a probabilistic/mechanistic analysis holds that 'A causes B' says something about probabilities/mechanisms respectively, while the proponent of the epistemic theory holds that it says something about rational belief.

It is not hard to see how this epistemic view of causality gets round the problems that beset probabilistic causality, mechanistic causality, and a pluralist combination of the two. The

counterexamples to probabilistic causality are not counterexamples to epistemic causality because PCC is not an assumption of epistemic causality. Under the epistemic view, PCC may hold in certain circumstances (Williamson 2005), but it is not guaranteed to hold: if in a particular case one should not posit a common cause to account for a probabilistic dependence, then the ideal causal epistemology, when applied to the ideal evidence set, will not yield a common cause. Similarly, under the epistemic view, cause and effect need not be probabilistically dependent. If, as in the example of sect.10, one should conclude that *B* causes *C* even though the two are probabilistically independent, then by construction of the epistemic theory, it will. Thus problems with probabilistic causality do not carry over to epistemic causality. Turning to difficulties with the mechanistic approach, we see that absences are no problem for the epistemic theory. Since the link between cause and effect is not physical, causes and effects need not be physical entities either. Further, if one should not conclude from the fact that there are certain mechanisms linking a beating heart and a failing liver that the former is causing the latter, then by construction the epistemic theory won't. Heterogeneity of mechanisms across the sciences is no problem because the causal relation is not analysed in terms of those mechanisms but in terms of rational belief, an account that is not specific to particular sciences. So problems with the mechanistic theory are not problems for the epistemic theory. Moving on to pluralism, it is clear that the epistemic approach is not pluralist, so it can account for the homogeneity of causal talk. Finally, the epistemological problems with pluralism and the two monistic accounts cannot carry over to epistemic causality. If the claim that smoking causes cancer requires both probabilistic and mechanistic evidence then it requires both kinds of evidence in the ideal causal epistemology, and hence under the epistemic account.

Of course the epistemic theory of causality may be subject to problems of its own. Here we shall consider only a few possible objections; further discussion can be found in Williamson (2005; 2006*a*; 2006*b*; 2007); Choi (2006). First, if in the ideal causal epistemology causal beliefs are constrained by known mechanisms as well as probabilistic dependencies, and mechanisms are themselves causal, then won't the epistemic account offer an account of causality in terms of causality—a circular account?[7] Of course it is a matter of debate as to whether mechanisms—or more generally whatever does the explanatory work in science—are themselves causal. Russell (1913) argued that at base science is not causal; others disagree. But we do not need to decide this question here, because even if mechanisms are causal, that does not mean that the epistemic account is viciously circular. I don't dispute that in the ideal causal epistemology evidence of mechanisms as well as dependencies helps to constrain appropriate causal beliefs. But this constraining relation is epistemological rather than ontological: if you grant certain facts, then certain beliefs are appropriate. For there to be any vicious circularity, it would have to be the case that you could not grant those facts without having the beliefs in the first place. That is, there would have to be an epistemological circularity. But there is no such circularity: we know about mechanisms linking a beating heart to a failing liver without needing to know whether or not the former causes the latter. Perhaps, in order to know about this mechanism one needs to know about other causal relationships—lower-level relationships that concern the circulation of blood. But even if that were the case, there would be no vicious circularity because these other causal relationships are not in question here: it is the higher-level relationship that is in question. There is no

reason why the ideal causal epistemology should not involve feedback—that is, certain evidence warrants certain causal beliefs which in turn allow the prospect of further evidence which then warrants other causal beliefs. It is only a problem if one already needs to believe that *A* causes *B* in order to believe that *A* causes *B*; but that is not the case here.

The preceding response applies equally to a second objection. Causal relations can themselves be causes and effects (Williamson and Gabbay, 2005). For instance, smoking causing cancer causes governments to restrict tobacco advertising. Again, there is a whiff of circularity here: do we not need to know causal relations in order to determine causal relations? But again, there is no vicious circularity. While it seems reasonable to hold that one needs to know whether smoking causes cancer before one can tell whether or not that causal relationship is a cause of advertising restrictions, the former causal relation is at a lower level to the latter. As before, the causal epistemology is incremental—some causal beliefs are required before others can be acquired. It is not circular in the sense that a particular causal belief presupposes itself.

There is a third related objection. Under the above account there is a category difference between causal relationships on the one hand, which are epistemic, and probabilistic dependencies and mechanisms on the other, which are physical. But what is the basis for this sharp distinction? If there is no substance to it and dependencies and mechanisms turn out to be epistemic, then the concept of ideal evidence would be difficult to delineate. In response I would say this: I have argued that one needs to plump for epistemic causality because of the failure of analyses in terms of probabilities or mechanisms, or a pluralist combination of the two. Purely on grounds of simplicity it would be very nice to have some such analysis, but unfortunately it is not possible, and we are forced to turn to an epistemic account. I have not argued for epistemic causality on the basis of general considerations in favour of epistemic accounts or general considerations that tell against physical accounts. Thus unless there are compelling reasons why one can't construe evidence of dependencies and evidence of mechanisms in terms of physical probabilities and physical processes respectively, it is quite natural to maintain a sharp distinction between epistemic and physical entities. Now I grant that there are problems with physical accounts of probability—it is implausible that a physical interpretation of probability will be successful in underpinning all our uses of probability. For example, one may rightly say that there is probability 0.1 that the trillionth digit of p is 9, but a physical account is unlikely to make sense of this claim. However, there seems no problem construing probabilistic *evidence* in terms of physical probabilities—the dependence between smoking and cancer for instance can quite easily be interpreted in terms of relative frequencies. Similarly there does not seem to be any insurmountable problem with construing evidence of the mechanism from smoking to cancer in terms of the physical features of the human body and of smoke. Hence there do not seem to be any compelling reasons for abandoning physical accounts of probabilistic and mechanistic evidence, and the distinction between epistemic and physical can be maintained.

A fourth related objection proceeds as follows. One may grant that both probabilities and mechanisms are physical, but claim that there is no sharp ontological distinction between probabilistic dependencies and mechanisms. (Perhaps on the grounds that mechanisms are ultimately reducible to low-level chains of probabilistic dependencies; perhaps because probabilistic dependencies are somehow analysable in terms of mechanisms; or, more

plausibly, because both are analysable in terms of the basic make-up of the physical universe.) If probabilities and mechanisms are at base of the same stuff, and probabilistic dependencies and mechanisms support causal claims, then surely causality is of this stuff too—that is, causality is itself at base physical. In response, I would say that this objection is sound but misplaced. It is sound in the sense that indeed the epistemic theory makes no commitment to non-physical entities in order to analyse causality. According to the epistemic theory, the causal relation is determined by its uses—inference and explanation; causality is a map of optimal causal beliefs, and these beliefs are optimal in the sense that they chart the optimal inferences and explanations. Now it must be the physical world that makes these inferences and explanations optimal, so the physical world is at base the truthmaker for causal claims. The objection is misplaced, however, because the job of a philosophical theory of X is to do more than say what kind of thing the truthmakers of X-claims are *at base*. It is easy to say that fundamental physical entities and their spatio-temporal locations make X-claims true; it is of course much harder to say *how* they make them true. If Zs make X-claims true but only via Ys, then a theory of X should point this out.[8] According to the epistemic account the physical world makes causal claims true, but only via probabilistic dependencies, mechanisms, and rational beliefs. Some maps directly map the world, others map inferences—as the epistemic theory makes clear, causality is a map of the latter kind.

This point can be put another way. If the epistemic account of causality is right, it must in principle be possible to come up with some characterization of causality that just appeals to the indicators of causality, along the lines of 'C causes E iff there is a dependency and a mechanism and C is prior to E unless C or E are absences in which case ...' However, the resulting characterization would be so complicated that it would be very hard to see why it is a correct characterization and why we should have a concept of cause at all. The answer to these questions, I think, must invoke the uses of the causal relation and the idea of inferential and explanatory success. So the epistemic account is to be preferred on the grounds (1) that it is clear that its characterization of causality must be correct and (2) that it tells us the full story, while the above kind of characterization is just a part of the picture.

Here's an analogy. Consider a travel graph whose nodes are towns and where one node is linked to another by an edge if normally you can travel between them in two hours. This kind of graph allows one to make a whole host of useful inferences and explanations related to travelling. Let's call the binary relation that is depicted by the edges of this map 'travelity'. Now one could try to characterize this relation in terms of its evidential indicators as follows 'C travelizes E iff there is some kind of mechanism for travelling between C and E for which the mean travel time is less than two hours unless it is a Sunday or bank holiday or ...' Already it's getting a bit pointless. It's just a map. It's a very useful kind of map precisely because it over-simplifies and because it overloads a simple binary relation with connotations of travel mechanisms (which are explanatory) and travel times (which enable inferences). While travelity may supervene on physical entities and their spatio-temporal locations, any viable account of travelity should talk a bit about the map and its uses.

I have argued in this chapter that probabilistic theories of causality are inadequate in two respects: they admit counterexamples and they fail to account for the relationship between causality and mechanisms. But mechanistic theories are no better, nor are pluralist accounts

that are part probabilistic, part mechanistic. The right way to integrate probabilistic and mechanistic considerations into an account of causality is to embed them in an epistemic account.

In terms of the distinctions of sect. 2, the epistemic account can cover both single-case and generic causal claims, since we can have both single-case and generic causal beliefs. Similarly causes and effects can be population-level or individual-level. Causality turns out not to be directly physical on the epistemic account, since it is analysed in terms of the way we should represent the world, rather than directly in terms of the world itself. The epistemic account can accommodate an objective notion of cause or indeed a subjective notion, though I have suggested in Williamson (2005: §9.7) that it is likely that ideal causal beliefs are so highly constrained that they are fully objective or close to fully objective. The epistemic account can handle both potential and actual causation; the former is predominantly associated with the inferential uses of causality while the latter is primarily used for explanation.

FURTHER READING

A good grounding in probabilistic causality can be had by studying the following texts. Reichenbach (1971) is a readable and historically important first port of call (see sect. 4). Salmon (1988) gives a recent exposition of his own view of causality: while Reichenbach gives primacy to probabilistic relationships over mechanisms in his account, Salmon does the opposite. Pearl (2000) offers a comprehensive picture of the causal net approach outlined in sect. 7. The motivation behind the epistemic theory of causality is presented in Williamson (2005). Hitchcock (1997) is also recommended as a survey of probabilistic causation that covers several issues not taken up here.

REFERENCES

BAUER, M. (1998). 'The Medicalization of Science News—from the "Rocket-Scalpel" to the "Gene-Meteorite" Complex', *Social Science Information* 37/4: 731–51.

CARTWRIGHT, N. (1979). 'Causal Laws and Effective Strategies', *Noûs* 13: 419–37.

CHOI, S. (2006). Review of 'Bayesian Nets and Causality', *Mind* 115: 502–6.

DOWE, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

EELLS, E. (1991). *Probabilistic Causality*. Cambridge: Cambridge University Press.

—— and SOBER, E. (1983). 'Probabilistic Causality and the Question of Transitivity', *Philosophy of Science* 50: 35–57.

GILLIES, D. (2000). *Philosophical Theories of Probability*. London: Routledge.

GOOD, I. J. (1959). 'A Theory of Causality', *British Journal for the Philosophy of Science* 9: 307–10.

—— (1961*a*). 'A Causal Calculus I', *British Journal for the Philosophy of Science* 11: 305–18. Errata ibid. 13: 88.

—— (1961*b*). 'A Causal Calculus II', *British Journal for the Philosophy of Science* 12: 43–51. Errata ibid. 13: 88.

HALL, N. (2004). 'Two Concepts of Causation'. in J. Collins, N. Hall, and L. Paul (eds.),

Causation and Counterfactuals. Cambridge, Mass.: MIT, 225–76.

HITCHCOCK, C. (1997). 'Probabilistic Causation' *Stanford Encyclopedia of Philosophy*. Rev. edn. 2002.

MACH, E. (1883). *The Science of Mechanics*. Stanford, Calif.: University of Stanford Press. 4th edn. 1919. La Salke, Ill.: Open Court.

—— (1905). *Knowledge and Error*. Dordrecht: Reidel.

MARTEL, I. (2000). '*Probabilistic Empiricism: In Defence of a Reichenbachian Theory of Causation and the Direction of Time*'. PhD thesis, University of Colorado.

NEAPOLITAN, R. E. (1990). *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. New York: Wiley.

OTTE, R. (1986). 'Reichenbach, Causation, and Explanation', in *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, i. 59–65.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, Calif: Morgan Kaufmann.

—— (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

PEARSON, K. (1892). *The Grammar of Science*. 2nd edn. 1900. London: Black. 3rd edn. 1911. New York: Macmillan.

REICHENBACH, H. (1959). 'The Principle of Causality and the Possibility of its Empirical Confirmation', in *Modern Philosophy of Science*. London: Routledge & Kegan Paul, 109–34.

—— (1971). *The Direction of Time*. Berkeley and Los Angeles: University of California Press.

REISS, J. (2007). 'Time Series, Nonsense Correlations and the Principle of the Common Cause', in *Causality and Probability in the Science*. London: College Publications, 179–96.

RUSSELL, B. (1913). 'On the Notion of Cause', *Proceedings of the Aristotelian Society*, 13: 1–26.

RUSSO, F., and WILLIAMSON, J. (2007). 'Interpreting Causality in the Health Sciences', *International Studies in the Philosophy of Science* 21/2: 157–170.

SALMON, W. C. (1980). 'Probabilistic Causality', *Pacific Philosophical Quarterly* 61: 50–74; repr. in Salmon (1998), 208–32.

—— (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

—— (1988). 'Intuitions—Good and Not-so-good', In B. Skyrms and W. L. Harper (eds.), *Causation, Chance, and Credence* Dordrecht: Kluwer, i. 51–71.

—— (1998). *Causality and Explanation*. Oxford: Oxford University Press.

SKYRMS, B. (1980). *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven, Conn.: Yale University Press.

SOBER, E. (1988). 'The Principle of the Common Cause', in J. H. Fetzer (ed.), *Probability and Causality: Essays in Honour of Wesley C. Salmon*. Dordrecht: Reidel, 211–28.

—— (2001). 'Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause', *British Journal for the Philosophy of Science* 52: 331–46.

SPIRTES, P., GLYMOUR, C., and SCHEINES, R. (1993). *Causation, Prediction, and Search*.

2nd edn. 2000. Cambridge Mass.: MIT.

Spohn, W. (2002). 'Bayesian Nets Are All There Is to Causal Dependence', in M. C. Galavotti, P. Suppes, and D. Costantini (eds.), *Stochastic Causality*. Chicago: University of Chicago Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Williamson, J. (2005). *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford: Oxford University Press.

—— (2006*a*). 'Causal Pluralism versus Epistemic Causality', *Philosophica* 77: 69–96.

—— (2006*b*). 'Dispositional versus epistemic causality', *Minds and Machines,* 16: 259–76.

—— (2007). 'Causality', in D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*. Berlin: Springer, xiv. 95–126.

—— and Gabbay, D. (2005). 'Recursive Causality in Bayesian Networks and Self-fibring Networks', in D. Gillies (ed.), *Laws and Models in the Sciences*. London: King's College Publications, 173–221. With comments, 223–45.

Yule, G. U. (1926). 'Why Do We Sometimes Get Nonsense-Correlations Between Time Series?' A Study in Sampling and the Nature of Time Series, *Journal of the Royal Statistical Society*, 89/1: 1–63.

CHAPTER 10
CAUSAL PROCESS THEORIES

PHIL DOWE

1. INTRODUCTION

If the core idea of process theories of causation is that causation can be understood in terms of causal processes and interactions, then the approach should be attributed primarily to Wesley Salmon (1925–2001). Salmon takes causal processes and interactions as more fundamental than causal relations between events. To express this Salmon liked to quote John Venn: 'Substitute for the time honoured "chain of causation", so often introduced into discussions upon this subject, the phrase a "rope of causation", and see what a very different aspect the question will wear' (Venn 1866: 320). According to the process theory, any facts about causation as a relation between events obtain only on account of more basic facts about causal processes and interactions. Causal processes are the world-lines of objects, exhibiting some characteristic essential for causation.

There are other approaches to causation that seek to respect the idea of a continuous process in terms of chains of events linked by the appropriate relations (e.g. Sober 1988; Menzies 1989; Hitchcock 2001; Schaffer 2001; Thalos 2002). Since the 'appropriate relations' tend to be counterfactual dependence, chance raising, or lawful sequence, these accounts are best viewed as instances of other theories of causation dealt with in this volume (Ch. 7 on regularity theories, Ch. 8 on counterfactual theories, Ch. 9 on probabilistic theories, and Ch. 14 on causal modelling).

A closer relative to process theories is the class of theories that takes causation to be the transfer or persistence of properties of a specific sort (e.g. Fair 1979; Ehring 1997). Some examples of this important class of theories are summarized in the final section. Another close relative is the approach that focuses on mechanisms (e.g. Machamer, Darden, and Craver 2000). Accounts of mechanisms are discussed in Ch. 15 of this volume.

2. THE SALMON PROGRAMME

While many of Salmon's ideas on causation carry some debt or other to his Ph.D. supervisor Hans Reichenbach (we will not trace those debts in this chapter), the theory of causal processes and interactions is itself original to Salmon.

Salmon's interest in causation arose from his work on scientific explanation where, in response to Hempel, Salmon argued for a causal theory of explanation (e.g. Salmon 1978).

This of course begs a theory of causation. However, Salmon wanted to avoid the shortcomings of other popular theories of causation. First, there are general problems with basing causation on events. Further, Salmon felt the modal commitments of counterfactual theories were a violation of actualist requirements of empiricism, that the probabilistic theories of Reichenbach and Good faced unsurmountable technical difficulties (Salmon 1998: ch. 14), and that any account that ultimately appeals to agency misrepresents the fundamental objective and 'ontic' nature of causation. Salmon's views were further shaped by his work on space and time where he was impressed by the idea that relativity could be formulated in terms of processes rather than events, and by his interest in the causal theory of time.

3. PROBLEMS WITH EVENTS

The problem with events was raised by Bertrand Russell in his paper 'On the Notion of "Cause"'. Famously, Russell observes 'The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm' (Russell [1913] 1957: 174). Russell argued that the philosopher's concept of causation involving, as it does, the law of universal determinism that every event has a cause and the associated concept of causation as a relation between events, is 'otiose' and replaced in modern science by the concept of causal laws understood in terms of functional relations, where these causal laws are not necessarily deterministic. Russell (ibid. 177) was concerned with the law of causality in the form: 'Given any event e_1, there is an event e_2 and a time interval τ such that, whenever e_1 occurs, e_2 follows after an interval τ'. First, Russell felt that because the time series is dense, no two times are contiguous, and τ must be of finite duration. Second, events may be defined more or less 'narrowly': 'Jenny's drinking arsenic' is fairly widely defined, 'Jenny's drinking arsenic together with her already having certain stomach contents together with her being on a hillside miles from civilization' is more narrowly defined.

A dilemma then arises. On the one hand, for an event to recur 'it must not be defined too narrowly' (ibid. 180). For example, the event must be something short of the whole state of the universe because it is very probable that such an event e_1 will never recur, in which case it becomes trivially true that event e_1 is the cause of an event e_2 which follows it by the suitable time interval τ, since the relevant regularity has just the one instance. On the other hand, if we define the cause 'sufficiently widely' to allow significant recurrence, then since the time interval between the cause and the effect is of finite duration it is possible that in the intervening interval something will happen that prevents the effect.

For example, Jenny's drinking arsenic caused her death. However, drinking arsenic will only be followed by death if one does not have a stomach full of alkaline, and if there is no stomach pump on hand, and so on. But if we define the cause of death as 'drinking arsenic in the absence of large amounts of alkaline and stomach pumps, etc.' then given the likely extent of the 'etc.' this narrowly defined event is unlikely to recur.

There's another problem with going narrow (which, because his interest focuses on the law of causality, Russell doesn't mention): in the case that intuitively, and on a wide definition

event e_3 (not e_1) is the cause of e_2, then the theory may wrongly say that e_1 is the cause of e_2, if the narrow definition of e_1 includes e_3, and so defined e_1 is always followed by e_2. If Jenny takes chocolate with the arsenic then her chocolate eating is the cause of her death if all cases of chocolate plus arsenic consumption are followed by deaths.

Of course one might hope that there is some middle ground, a definition of the cause sufficiently wide that it recurs but sufficiently narrow that it will always be followed by its effect. Russell simply makes the point that mature science is not concerned with identifying such, but rather focuses on functional relations between variables, a focus far removed from the kind of common-sense events that we take to be causes and effects. Salmon, however, felt that these problems could be circumvented by focusing on spatio-temporally continuous processes.

4. CAUSAL LINES AND PSEUDO PROCESSES

An important forerunner of Salmon's account of causal processes is Bertrand Russell's account of causal lines in his *Human Knowledge* (1948). There Russell elaborates the view that 'causal lines' replace the primitive notion of causation in the scientific view of the world, and not only replace but also explain the extent to which the primitive notion, causation, is correct. He writes, 'When two events belong to one causal line the earlier may be said to "cause" the later. In this way laws of the form "A causes" may preserve a certain validity' (1948: 334). And

I call a series of events a 'causal line' if, given some of them, we can infer something about the others without having to know anything about the environment. (ibid. 333)

A causal line may always be regarded as a persistence of something, a person, a table, a photon, or what not. Throughout a given causal line, there may be constancy of quality, constancy of structure, or gradual changes in either, but not sudden change of any considerable magnitude. (ibid. 475–7)

So the trajectory through time of something is a causal line if it doesn't change too much, and if it persists in isolation from other things. A series of events with this kind of similarity displays what Russell calls 'quasi-permanence'.

Although this retains a focus on relations between events, Salmon was nevertheless interested in the idea of a causal line with its 'constancy of structure'. However Salmon felt that Russell's theory of a causal line does not enable an important distinction between pseudo and causal processes to be made. As Reichenbach (1958: 147–9) argued, as he reflected on the implications of Einstein's special theory of relativity, science requires that we distinguish between causal and pseudo processes. Reichenbach noticed that the central principle that nothing travels faster than the speed of light is 'violated' by certain processes. For example, a

spot of light moving along a wall is capable of moving faster than the speed of light. (One needs just a powerful enough light and a wall sufficiently large and sufficiently distant.) Other examples include shadows and the point of intersection of two rulers (see Salmon's clear exposition in his 1984: 141–4). Such pseudo processes, as we shall call them (Reichenbach called them 'unreal sequences', 1958: 147–9), do not violate special relativity, Reichenbach argued, simply because they are not causal processes, and the principle that nothing travels faster than the speed of light applies only to causal processes. Thus special relativity demands a distinction between causal and pseudo processes. But Russell's theory doesn't explain this distinction, because both causal processes and pseudo processes display constancy of structure and quality; and both licence inferences of the sort Russell has in mind. For example, the phase velocity of a wave packet is a pseudo process but the group velocity is a causal process; yet both licence reliable predictions.

5. Causal Processes and Interactions: The Mark Criterion

We now turn to Salmon's positive account of causation proposed systematically in Salmon (1984). For Salmon the causal structure of the world consists in the nexus of causal processes and interactions. A process is anything with constancy of structure over time. These may be either causal or psuedo; a causal process is a process capable of transmitting a mark. A 'mark' is any local modification of a characteristic, while 'transmission' is understood in terms of the 'at-at' theory. The latter is expressed in the principle of mark transmission (MT), which states:

MT: Let P be a process that, in the absence of interactions with other processes would remain uniform with respect to a characteristic Q, which it would manifest consistently over an interval that includes both of the spacetime points A and B (A ≠ B). Then, a mark (consisting of a modification of Q into Q*), which has been introduced into process P by means of a single local interaction at a point A, is transmitted to point B if [and only if] P manifests the modification Q* at B and at all stages of the process between A and B without additional interactions. (ibid. 148)

Salmon himself omits the 'only if' condition; reasons to include it have been given by Sober (1987: 253) and Dowe (1992: 198).

For example, a ball moving through the air is a causal process, since it can be marked, for example by making a small cut with a knife; and this mark would be transmitted since the cut would continue to exist at later times on the ball's trajectory provided there is no further interaction. On the other hand, a spot of light moving across a wall does not transmit a mark due to a single local modification. For example a change in the shape of the light spot caused by distorting the surface is not subsequently transmitted and so doesn't count as a mark transmission.

A causal interaction involves the mutual modification of two intersecting processes:

CI: Let P_1 and P_2 be two processes that intersect with one another at the spacetime point S, which belongs to the histories of both. Let Q be a characteristic of that process P_1 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P_1) if the intersection with P_2 did not occur; let R be a characteristic that process P_2 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P_2) if the intersection with P_1 did not occur. Then, the intersection of P_1 and P_2 at S constitutes a causal interaction if (1) P_1 exhibits the characteristic Q before S, but it exhibits a modified characteristic Q' throughout an interval immediately following S; and (2) P_2 exhibits R before S but it exhibits a modified characteristic R' throughout an interval immediately following S. (Salmon 1984: 171)

For example, the collision of two balls is a causal interaction when both balls undergo a change in momentum, since both balls would have continued with their original momenta had the collision not occurred.

Both MT and CI involve counterfactuals. The former defines a causal process in terms of possible marks, and further, the definition of mark transmission requires that the changed characteristic would otherwise have remained constant. Salmon was uncomfortable with the use of counterfactuals, fearing that their context dependence was problematic for an 'ontic' account. He noted, however, that the truth of these counterfactuals should be determinable from scientific experiments, and would, one may speculate, have found support for this idea in the account of Woodward (2003). However, as Kitcher notes, this undermines the idea that the process theory is a distinct approach to causation:

I suggest that we can have causation without linking causal processes. What is critical to the causal claims seems to be the truth of the counterfactuals, not the existence of the processes and the interactions. If this is correct then it is not just that Salmon's account of the causal structure of the world needs supplementing through the introduction of more counterfactuals. The counterfactuals are the heart of the theory, while the claims about the existence of processes and interactions are, in principle, dispensable. Perhaps these notions may prove useful in protecting a basically counterfactual theory of causation against certain familiar forms of difficulty (problems of pre-emption, overdetermination, epiphenomena, and so forth). But, instead of viewing Salmon's account as based on his explications of process and interaction, it might be more revealing to see him as developing a particular kind of counterfactual theory of causation, one that has some extra machinery for avoiding the usual difficulties that beset such proposals. (Kitcher 1989: 472; see also Psillos 2002: 112–18)

The MT account has other problems that ultimately led Salmon himself to abandon it. Criticism focused on the notion of a characteristic, which would seem to be underspecified. For example, the definition MT allows pseudo processes such as shadows to count as causal. A shadow having the property of 'occurring after a certain time' (Sober 1987: 254), of 'being the shadow of a scratched car' (Kitcher 1989: 638) or of 'being closer to the Harbour Bridge than

to the Opera House' (Dowe 1992: 201) will qualify as a causal process under MT.

Further, there are cases of 'derivative marks' (Kitcher 1989: 463) where a pseudo process displays a modification in a characteristic on account of a change in the causal processes on which it depends, either in the source or in the causal background. A change at the source would include cases where the spotlight spot is 'marked' by a coloured filter at the source (Salmon 1984: 142) or a car's shadow is marked when a passenger's arm holds up a flag (Kitcher 1989: 463). Salmon's clause 'by means of a *single* local interaction' (MT) is intended to exclude this type of case. But this will not always work: take the case where a stationary car throws its shadow on a fence. Suddenly the fence falls over, producing a permanent modification in the shadow. Then the shadow has been marked by the single local action of the falling fence (Dowe 1992: 201–2).

Criticisms such as these, together with his discomfort with counterfactuals led Salmon to abandon the mark criterion in favour of the conserved quantity theory, to which we now turn.

6. Conserved Quantity Theory

If causal processes are causal in virtue of *actual* features, rather than counterfactual features, what would those features be? They should be general features common to all causal processes, but lacking in pseudo processes. The result in relativity theory that only time-like world-lines transmit energy suggests the answer might be the transmission of energy. The link between energy and causation is not new. It was suggested by Quine (1973: 5); and Fair and Aronson both formulated theories of causation in terms of energy transfer. And in any case the idea was common enough in the eighteenth and nineteenth centuries. (Krajewski (1997: 194–5) traces the history from Mayer through Lorenz.) The idea of the conserved quantity theory is to generalize this notion to all types of physical interactions. (Brian Skyrms (1980: 111) first suggested 'conserved quantities' and Dowe (1992) formulated this as a Salmon-style process theory.) Dowe (1992; 1995; 2000) and Salmon (1994; 1997) have offered various versions; the most recent are:

CQ1. A *causal interaction* is an intersection of world lines that involves exchange of a conserved quantity.

CQ2. A *causal process* is a world line of an object that possesses a conserved quantity. (Dowe 1995: 323)

Definition 1. A causal interaction is an intersection of world-lines that involves exchange of a conserved quantity.

Definition 2. A causal process is a world-line of an object that transmits a nonzero amount of a conserved quantity at each moment of its history (each spacetime point of its trajectory).

Definition 3. A process transmits a conserved quantity between A and B (A ≠ B) if it possesses [a fixed amount of] this quantity at A and at B and at every stage of the process between A and B without any interactions in the open interval (A, B) that involve an exchange

of that particular conserved quantity.   (Salmon 1997: 462, 468)

A conserved quantity is any quantity governed by a conservation law. To state the obvious, current scientific theory is our best guide as to what these are: quantities such as mass-energy, linear momentum, and charge. Concerns have been raised about the appeal here to conservation laws. It is common to define conservation in terms of constancy within a closed system. As Hitchcock (1995: 315–16) points out, it would be circular to define a 'closed system' as one that is not involved in causal interactions with anything external. Dowe (2000: 95) suggests 'we need to explicate the notion of a closed system in terms only of the quantities concerned. For example, energy is conserved in chemical reactions, on the assumption that there is no net flow of energy into or out of the system.' Schaffer (2001: 810) comments that this 'looks to invoke the very notion of "flow" that the process account is supposed to analyze'. McDaniel (2002: 261) suggests two possible responses to this. First, the theory could simply list the quantities held to be relevant to causation, Second, the theory could appeal directly to universally conserved quantities, in other words, doing away with appeal to any closed system besides the universe itself. This may run up against the objection of Reuger discussed below.

However, Sungho Choi (2003) has provided a thorough examination of possible definitions of a closed system, and proposes:
DC: A system is closed with respect to a physical quantity Q at a time t iff

$$\text{(a) } dQ_{in}/dt = dQ_{out}/dt = 0 \; attor, \text{ (b)} dQ_{in}/dt = -dQ_{out}/dt = 0 att$$

where Q_{in} is the amount of Q inside the system and is Q_{out} the amount of Q outside the system. (ibid. 519)

For vector quantities the definition must apply to all components of the vector. This, Choi argues, does not involve any circular appeal to causation.

Alexander Reuger (1998) has argued that since in some general relativistic spacetimes global conservation laws will not hold, it would seem to follow that in such a spacetime there would not be causal processes at all. Dowe's (2000: 97–8) response is that our world is not such a spacetime. (*Ad hominem*, this may be a particular problem for Dowe who argues elsewhere that time travel and hence causation is possible in such spacetimes. See Schaffer 2001: 811.)

John Norton, while endorsing the Salmon–Dowe tack of not tying the theory to any particular conserved quantity since that leaves the theory hostage to scientific developments, nevertheless warns that 'if we are permissive in selection of the conserved quantity, we risk trivialization by the construction of artificial conserved quantities specially tailored to make any chosen process come out as causal' (2007: 18–19).

The differences between Salmon and Dowe indicated above in definitions 1–3 focus

attention on the distinction between pseudo and causal processes. For Salmon it is important that the conserved quantity is transmitted, and indeed that a fixed quantity is transmitted in the absence of interactions, in order to rule out cases of 'accidental' process-like energy appearances. Dowe has concerns about the directionality built into 'transmission', and instead attempts to rule out accidental processes via the identity through time of the object in question. So, for Salmon the spotlight spot does not transmit energy in the absence of interactions, but involves a continual string of interactions. For Dowe it is not the spot that possesses energy, but rather the various distinct patches of wall illuminated.

Hitchcock (1995) produces this counterexample: consider an object casting a shadow on the surface of a charged plate. At each point of its trajectory the shadow 'possesses' a fixed charge. But shadows are the archetypical pseudo process. Dowe (2000: 98–9) and Salmon (1997: 472) claim that it is the plate that possesses the charge, and the shadow that moves. Salmon (ibid.), however, suggests that the more problematic 'object' is the series of plate segments currently in shadow, in Dowe's terminology a 'time-wise gerrymander'. Salmon's answer to this is that this object does not transmit charge or else charge in a region would augment when the shadow passes over it, and he proposes to add the corollary explicitly to apply the conservation law to this kind of case (critiqued in detail by Choi 2002: 110–14): 'When two or more processes possessing a given conserved quantity intersect (whether they interact or not), the amount of that quantity in the region of intersection must equal the sum of the separate quantities possessed by the processes thus intersecting' (Salmon 1997: 473). On the other hand, Dowe's answer is that the world-line of the moving shadow is the world-line of an object that does not possess charge, while the world-line of the segments of shadowed plate is not the world-line of an object. (But see McDaniel 2002: 260; Garcia-Encinas 2004.)

Sungho Choi (2002: 114–15) offers a further counterexample to Salmon's version. Suppose the plate contains a boundary such that there is twice as much charge density on one side compared to the other. Suppose the shadow crosses from the lower density to the higher density. Consider the world-lines of (1) the gerrymandered objects which are the segments of plate when crossed by the shadow and (2) the segment of plate just before the boundary. Their intersection will count as a causal interaction on Salmon's account since the world-line in (1) exhibits a change in the conserved quantity.

7. MISCONNECTIONS: THE PROBLEM OF CAUSAL RELEVANCE

So far we have focused on the ability of the conserved quantity theory to distinguish causal processes from pseudo processes. But in the words of Chris Hitchcock (2004), *what's this distinction good for?* Salmon and Dowe claim that they are offering a theory of causation, yet each acknowledges one way or another that the definitions above at best gives just a necessary condition for two events to be related as cause and effect. As Woodward (2003: 357) notes, '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, putting a chalk mark on the white ball is a causal interaction linked by causal processes and interactions to the black ball's sinking (after the white ball strikes the black ball), yet it doesn't cause the black ball's sinking (ibid. 351).

Dowe offers the following account (restricting the causal relata to facts for simplicity):

Causal Connection: There is a causal connection (or thread) between a fact q(a) and a fact q′(b) if and only if there is a set of causal processes and interactions between q(a) and q′(b) such that:

(1) any change of object from a to b and any change of conserved quantity from q to q′ occur at a causal interaction involving the following changes: ⊿q(a), ⊿q(b), ⊿q′(a), and ⊿q′(a); and

(2) for any exchange in (1) involving more than one conserved quantity, the changes in quantities are governed by a single law of nature.

... where a and b are objects and q and q′ are conserved quantities possessed by those objects respectively. (Dowe 2000: sect. 7.4; see Hausman 2002: 720–1 for discussion)

The analysis would need to be expressed in a more general form for cases where there are more than two objects involved along the nexus of causal processes and interactions.

Condition (2) in the definition of causal connection states 'for any exchange in (1) involving more than one conserved quantity, the changes in quantities are governed by a single law of nature'. This is an attempt to rule out accidentally coincident causal interactions of the sort identified by Miguel and Paruelo (2002). In one of their examples two billiard balls collide, and at the same instant, one of them emits an alpha particle. Condition (2) would not work for the case also mentioned by Miguel and Paruelo where the same quantity is exchanged in both interactions.

The account, if successful, tells us when two events are related causally, either as cause and effect or vice versa, or as common effects or causes of some event. It will not tell us which of these is the case (Hausman 2002: 719; Ehring 2003: 531–2). To do that, both Salmon and Dowe (2000: ch. 8) appeal to a Reich-enbachian fork asymmetry theory (Dowe's particular version of the latter has been subject to serious critique by Hausman (2002: 722–3), which includes the point that his account of priority has nothing to do with the conserved quantity theory).

Suppose a rolling steel ball is charged at a certain point along its trajectory. Suppose its trajectory is unaffected, and the ball subsequently hits another ball. The account should tell us that the fact that the ball gets charged is not causally relevant to the fact that it hits the second ball. It does, since although on the Salmon–Dowe theory the ball's rolling is a causal process and the charging and the collision are causal interactions, and further, a change in the ball's charge and the change in the ball's momentum are both the kinds of changes envisaged in (1), nevertheless there is no causal interaction linking the ball's having charge to the ball's having momentum as required in (1). Hence there is no causal thread as defined in (1) linking the two facts.

The account should also tell us that the tennis ball's heading towards the wall is not the cause of the wall's being stationary after the ball bounces off. It does, because although there

is a set of causal processes and interactions linking these two events, there is a change of object along the 'thread'—ball to wall—yet the wall undergoes no change in momentum, which it needs for the set of causal processes and interactions to count as a causal connection on this definition. (But cf. Hausman 2002: 721; Twardy 2001: 268)

One might hope that the theory also tells us that the fact that a chalk mark is put on the white ball is not causally relevant to the fact that the black ball sinks since there is no causal thread as defined in (1) linking those two facts. However, such a results awaits a translation of 'chalking a ball' to a state involving a conserved quantity. (See the following section for a discussion of this issue.)

To this account Dowe (2000: sect. 7.4) adds the restriction that the facts that enter into causation should not be disjunctive. This is meant to deal with the following type of example. Suppose 'in a cold place, the heater is turned on for an hour, bringing the room to a bearable temperature. But an hour later the temperature is unbearable again, say 2°C. Then ... the fact that the heater was turned on is the cause of the fact that the temperature is unbearable at the later time'. According to Dowe 'the temperature is unbearable' is a disjunctive fact, meaning 'the temperature is less than x' for a certain x, which in turn means 'the temperature is y or z or ... '. The effect is simply that the room is 2°C. According to Ehring (2003: 532) this result remains counterintuitive.

8. COMMON SENSE, SCIENCE, AND REDUCTION

The Conserved Quantity theory is claimed by both Salmon and Dowe to be an *empirical analysis*, by which they mean that it concerns an objective feature of the actual world, and that it draws its primary justification from our best scientific theories. 'Empirical analysis' is to be contrasted with *conceptual analysis*, the approach that says in offering a theory of causation we seek to give an account of the concept as revealed in the way we (i.e. folk) think and speak. Conceptual analysis respects as primary data intuitions about causation; empirical analysis has no such commitment (Dowe 2000: ch. 1).

This construal of the task of delivering an account of causation has drawn criticism from a number of commentators. According to Koons (2003: 244) it threatens 'to turn [the] metaphysical account into a watered-down version of more-or-less contemporary physical theory'. But Hausman (2002: 718) notes that since causation is not a technical concept in science, '[w]ithout some plausible connection to what ordinary people and scientists take to be causation, the conserved quantity theory would float free of both physics and philosophy' (see also Garcia-Encinas 2004: 45). And McDaniel (2002: 259) asks what could justify one in believing a putative 'empirical analysis'. He adds that if an empirical analysis is not at least extensionally equivalent (in the actual world) to the true conceptual analysis, then what would be the point?

Despite their denial of a primary need to respect common sense intuitions about the concept of causation, Salmon and Dowe do still want to say their account deals with everyday cases of causation. This again raises the question of translation. As Kim (2001: 242) puts it, there is the 'question of whether the [Dowe–Salmon] theory provides a way to "translate" causality understood in the [Dowe–Salmon] theory into ordinary causal talk and *vice versa*' (and see

especially Hausman 1998: 14–17; 2002: 719).

According to Dowe the relata in true 'manifest' (common sense) claims of causation must be translated to physical states of the sort discussed in the previous section ('object *a* has a value *q* of a conserved quantity') such that the manifest causal claim supervenes on some physical causation. Even for purely physical cases such as 'chalking the ball' this is a complicated matter, and it is not obvious that it can be carried through.

Even if this could be made to work in purely physical cases, there remain questions about mental causation, causation in history, and causation in other branches of science besides physics (Woodward 2003: 355–6; Machamer, Darden, and Craver 2000: 7; Cartwright 2004: 812). In any case, to suppose that the conserved quantity theory will deal with causation in other branches of science also requires commitment to a fairly thoroughgoing reductionism, since clearly there is nothing in economics or psychology that could pass for a conservation law.

An alternative to such reductionism is the view developed by Nancy Cartwright, which we might call causal pluralism (see Ch. 16 below). After rejecting the conserved quantity theory (along with a range of major theories of causation) as an account of a 'monolithic' causal concept, on the grounds that it cannot deal with cases in economics, Cartwright (2004: 814) summarizes her position: '1. There is a variety of different kinds of causal laws that operate in a variety of different ways and a variety of different kinds of causal questions that we can ask. 2. Each of these can have its own characteristic markers; but there are no interesting features that they all share in common' (see also Hausman 2002: 723).

9. DISCONNECTIONS

'I killed the plant by not watering it' (Beebee 2004). If this is a case of causation then process theories are in trouble, because neither my not watering nor whatever I did instead are connected by a physical process to the plant's death. The same is true for 'my failure to check the oil caused my engine to seize'. Cases of causation by omission, absence, preventing (i.e. causing to not happen) and double prevention (Hall 2004) all raise the same difficulty. If these are cases of causation then the process theory cannot be right (Hausman 1998: 15–16; Schaffer 2000, 2004).

There is a long tradition that asserts that such are indeed cases of causation. Lewis (1986: 189–93; 2004) is adamant and Schaffer (2000; 2004) presents a detailed case. Others have denied these are indeed cases of causation (Aronson 1971; Dowe 1999; 2000; 2001; 2004; Armstrong 2004; Beebee 2004). Some extend their account of causation, in ways that depart from their respective central theses, to include such cases (Fair 1979: 246–7; Ehring 1997: 125, 139; Lewis 2004). According to Hall (2004) and Persson (2002) these cases show that there are two concepts of causation. According to Rieber (2002: 63–4) the account of causation in terms of the transfer of properties can handle these cases by translating negatives into the actual positives that obtain.

Dowe and Armstrong hold that while such cases are not genuine causation, they count as a close relative, which Dowe variously calls causation* (1999; 2000) or 'quasi-causation' (2001; cf. Ehring 1997: 150–1). Persson (2002) coins the term 'fake causation'. This relation is

essentially a counterfactual about causation (see also Fair 1979: 246–7). While admitting Schaffer's (2000) point that there are cases of quasi-causation which by intuition clearly count as causation, Dowe asserts that there is also an intuition of difference—other cases of quasi-causation which intuitively are not causation (2001; see also Rieber 2002). For a detailed rebuttal of the intuition of difference see Schaffer (2004: 209–11) and, from a Davidsonian perspective, Hunt (2005). Further, Dowe attempts to explain why we might confuse causation with quasi-causation by appealing to the similar roles they play in explanation, decision-making and inference, and justifies this similarity on the grounds of the relation between causation and quasi-causation (again, quasi-causation is essentially possible causation). Armstrong (2004) points out that another reason we might confuse the two concepts is that in practice it is often difficult to distinguish the two.

Dowe (2001: 221; see also 2000: sect. 6.4) offers this account of quasi-causation:

Prevention: A prevented B if A occurred and B did not, and there occurred an *x* such that

(P1) there is a causal interaction between A and the process due to *x*, and
(P2) if A had not occurred, *x* would have caused B.

where A and B name positive events or facts, and *x* is a variable ranging over events and/or facts.

For example, bumping the table (A) prevented the ball going into the pocket (B) because there is an interaction between bumping the table and the trajectory of the ball (*x*), a causal interaction, and the true counterfactual 'without A, *x* would have caused B'.

One reason that the above is stated only as a sufficient condition is that there is a need to account for alternative preventers, of which there are two types, preemptive prevention (cf. pre-emption) and overprevention (cf. overdetermination), since in both cases (P3) fails. To deal with the latter, Dowe (2000 sect. 6.4, adapted) disjoins (P2) with

(P2′) there exists a C such that had neither A nor C occurred, *x* would have caused B or ...

Suppose as well as bumping the table I subsequently knocked the moving ball with my elbow (C), again, preventing it from sinking (overprevention). (P2) is false, but by (P2′) A counts as a quasi-cause of B. So too does C, since substituted for A, it satisfies P(1). Suppose on the other hand C is some completely irrelevant event, and (P1–2) hold for A and B. Then although (P2′) holds for this A–C pair C will not count as a preventer of B because it does not satisfy (P1). (For a contrary view see Koons 2003: 246.)

Although the account in Dowe (2000) is unclear on this point, (P2′) will not handle pre-emptive prevention. Suppose I bumped the table, but didn't hit the ball with my elbow, although I would have had I not bumped the table. We need to add the further alternative:

(P2″) had A not occurred, *C* would have occurred and would have prevented *B*.

The possible prevention here is then analysed by (P1–2) from the perspective of that possible world.

Quasi-causation by omissions or absences are analysed thus:

Omission: not-A quasi-caused B if B occurred and A did not, and there occurred an *x* such that

(O1) *x* caused B, and
(O2) if A had occurred then A would have prevented B by interacting with *x*

where A and B name positive events/facts and *x* is a variable ranging over facts or events, and where prevention is analysed as above. (Dowe 2001: 222; see also 2000: sect. 6.5)

For example, being careful not to bump the table (not-A) quasi-caused the ball to sink (B) because the trajectory of the ball (*x*) causes B and had the table been bumped that would have prevented B. Further cases can be added: prevention by omission, and prevention of prevention, prevention of prevention of prevention, etc. (see Dowe 2000: sect. 6.6). There is indeed a great deal of quasi-causation around, as Beebee (2004) has argued.

Schaffer (2001: 811) offers two criticisms of the counterfactual theory of quasi-causation. First, he argues Salmon's and Dowe's process theory of causation is, ironically, ill equipped to tell us what genuine causation is in these possible worlds (i.e. the worlds one might take to be the truthmakers of the counterfactuals in P2 and O2) since theirs is only an account of causation in the actual world, and worse, if one follows the semantics of Lewis to deal with the counterfactuals, it will probably turn out that our conservation laws don't hold in those possible worlds At the very least, Dowe's (2001: 221) stated view that 'it's BYO semantics of counterfactuals' is not satisfactory. (For further discussion of this problem see Persson 2002: 139–40.) And second, the account is semantically unstable, since as Dowe asserts quasi-causation plays the same role as causation for explanation, decision theory and inference, that relation is a better deserver of the role of best fitting causation's connotations than Salmon–Dowe's 'genuine causation' (Dowe 2000: 296 n. 13; 2001: 811–12).

10. RELATED THEORIES OF CAUSATION

There is an increasing number of accounts of causation that are close relatives of the Process Theory but that don't exactly fit the definition of a Process Theory given above. In this section we summarize some important theories that take causation to be the transfer or persistence of properties of a specific property, in particular, energy.

10.1 Aronson's Transference Theory

Aronson's (1971: 422) theory is presented in three propositions:

(1) In '*A* causes *B*', '*B*' designates a change in an object, a change which is an *unnatural* one.

(2) In '*A* causes *B*', at the time *B* occurs, the object that causes *B* is in contact with the object that undergoes the change.

(3) Prior to the time of the occurrence of *B*, the body that makes contact with the effect object possesses a *quantity* (e.g. velocity, momentum, kinetic energy, heat, etc.) which is transferred to the effect object (when contact is made) and manifested as *B*.

Proposition (1) refers to a distinction Aronson draws between natural and causal changes —causal changes are those that result from interactions with other bodies; natural changes are not causal, and come about according to the normal course of events, when things happen without outside interference. Thus internal changes, or developments, are not seen by Aronson as cases of causation. Proposition (2) is Hume's requirement that causation occurs only by contact, which rules out action at a distance. It also means that, strictly speaking, there is no indirect causation, where one thing causes another via some intermediate mechanism. All causation is direct causation.

Proposition (3) is the key notion in Aronson's theory. It appeals to the idea of a quantity, which is possessed by objects, and which may be possessed by different objects in turn, but which is always possessed by some object. The direction of transfer sets the direction of causation. For a critique of this theory see Earman (1976).

10.2 Fair's Transference Theory

David Fair, a student of David Lewis, offers an account of causation similar in many respects to that of Aronson. Fair (1979) makes the claim that physics has discovered the true nature of causation: what causation really is, is a transfer of energy and/or momentum. This discovery is an empirical matter, and the identity is contingent. Fair presents his account as a programme for a physicalist reduction of the everyday concept, and he doesn't claim to be able to offer a detailed account of the way energy transfer makes true the fact that, for example, John's anger caused him to hit Bill. A full account awaits, Fair says, a complete unified science (ibid. 236).

Fair's programme begins with the reduction of the causal relata found in ordinary language. Events, objects, facts, properties, and so forth need to be redescribed in terms of the objects of physics. Fair introduces 'A-objects' and 'B-objects', which manifest the right physical quantities, namely energy and momentum, and where the A-objects underlie the events, facts, or objects identified as causes in everyday talk, while the B-objects underlie those identified as effects. The physical quantities, energy and momentum, underlie the properties that are identified as causes or effects in everyday causal talk.

The physically specifiable relation between the A-objects and the B-objects is the transfer of energy and/or momentum. Fair sees that the key is to be able to identify the same energy and/or momentum manifested in the effect as was manifested in the cause. This is achieved by specifying closed systems associated with the appropriate objects. A system is closed when no gross energy and/or momentum flows into or out of it. Energy and/or momentum transfer occurs when there is a flow of energy from the A-object to the B-object, which will be given by the time rate of change of energy and/or momentum across the spatial surface separating the A-object and the B-object.

Fair's (1979: 236) reduction thus is:

A causes B iff there are physical redescriptions of A and B as some manifestation of energy or momentum or [as referring to] objects manifesting these, that is transferred, at least in part, from the A-objects to the B-objects.

For one extended critique of Fair's theory see Dowe (2000: ch. 3).

10.3 Ehring's Trope Persistence Theory

Douglas Ehring sets out a highly original theory of causation in his book *Causation and Persistence* (1997). Ehring takes the relata of causation to be tropes—that is, non-repeating property instances. Causal connections involve the persistence of such tropes, and also their fission (partial destruction) and fusion. Trope persistence is endurantist, that is to say, tropes wholly exist at every time they exist, and that a particular trope at one time is strictly identical to itself at other times. Since tropes do not change they avoid the well-known problem for edurantists of temporary intrinsics.

Actually Ehring's theory has two parts. 'Strong causal connection' concerns trope persistence, and this is a symmetric matter. Causal priority on the other hand involves broader considerations including counterfactuals. Here are Ehring's definitions (following the summary in Ehring 2004):

Strong Causal Connection: Tropes *P* and *Q*, are strongly causally connected if and only if:

   (1) *P* and *Q* are lawfully connected, and either

   (2) *P* is identical to *Q* or some part of *Q*, or *Q* is identical to *P* or some part of *P*, or

   (3) *P* and *Q* supervene on tropes *P'* and *Q'* which satisfy (1) and (2).

Causal Priority: Ehring employs counterfactuals to define a relation of 'being a condition of a causal connection', and then he uses this relation, together with the symmetrical relation of causal connection, to define causal direction (1997: 145, 146, 148, 149, 151, 179).

Putting these two together, we get:

Causation: Trope P at t causes trope Q at t' iff either

(A) P at t is strongly causally connected to Q at t' and P at t is causally prior to Q at t' *or*

(B) there is a set of properties $(R_1, ..., R_n)$ such that P is a cause of R_1, under clause (A), ..., and R_n is a cause of Q under clause (A).

Clause (B) is to allow for events connected by a chain of indirect causation. For discussion of Ehring's theory see Beebee (1998).

10.4 Other Theories

There are a number of notable and related theories of causation which space unfortunately forbids us to deal with in detail. The reader is encouraged to consult the references for details.

On Castañeda's (1980) transference theory of causation, 'causity', is the transmission of a physical element: energy, movement, charge. According to Bigelow, Ellis, and Pargetter (1988) causation is the action of forces (see also Bigelow and Pargetter 1990), while for Heathcote (1989) causation is an interaction (as defined by a suitable quantum field theory). Collier (1999) develops the notion that causation is the transfer of information. Krajewski (1997) outlines several causal concepts including transfer of energy and the transfer of information. Kistler (1998; 2006) develops the trope persistence view in terms of conserved quantities. Rieber (2002) provides a conceptual analysis of causation in terms of property acquisition and transfer, and also gives references to many historical figures who hold a similar view. Finally, Chakravartty (2005) defines causal processes as systems of continuously manifesting relations between objects with causal properties and concomitant dispositions.

Further Reading

The early views of Salmon on spatio-temporally continuous processes, and interaction are presented in an accessible form in his first book on the topic (1984). His later views are mostly collected in (1998), which should be supplemented by Salmon (1997). The Conserved Quantity version is first offered in Dowe (1992) and in an extended form in (2000). Discussions of the distinction between causal and pseudo processes can be found in Hitchcock (1995; 2004), Choi (2002), McDaniel (2002), and Garcia-Encinas (2004). Discussions of the use of conserved quantities can be found in Hitchcock (1995), Rueger (1998), Schaffer (2001), McDaniel (2002), Choi (2002; 2003), and Norton (2007).

The problem of misconnections and causal relevance in general, as it relates to the Process Theory, is discussed in Twardy (2001), Hausman (2002), Miguel and Paruelo (2002), Ehring (2003), Woodward (2003), and Hitchcock (2004).

Responses to the notion of an empirical analysis are developed in Hausman (1998; 2002), Kim (2001), McDaniel (2002), Koons (2003), and Garcia-Encinas (2004).

Questions about the 'problem of disconnections' or apparent causation involving negatives, and its significance for process theories, are discussed in Ehring (1997), Hausman (1998), Schaffer (2000; 2004), Rieber (2002), Persson (2002), Koons (2003), Beebee (2004), Armstrong (2004), and Hunt (2005). Dowe's account can be found in (2001).

References

ARMSTRONG, D. (2004). 'Going through the Open Door Again', in J. Collins, N. Hall, and L. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 445–58.

ARONSON, J. (1971). 'On the Grammar of "Cause"'. *Synthese* 22: 414–30.

BEEBEE, H. (1998). 'Douglas Ehring, *Causation & Persistence*', *British Journal for the Philosophy of Science* 49: 181–4.

—— (2004). 'Causing and Nothingness', in J. Collins, N. Hall, and L. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 291–308.

BIGELOW, J., and PARGETTER, R. (1990). *Science and Necessity*. Cambridge: Cambridge University Press.

—— ELLIS, B., and PARGETTER, R. (1988). 'Forces', *Philosophy of Science* 55: 614–30.

CARTWRIGHT, N. (2004). 'Causation: One Word, Many Things', *Philosophy of Science* 71: 805–19.

CASTAÑEDA, H. (1980). 'Causes, Energy and Constant Conjunctions', in P. van Inwagen (ed.), *Time and Cause*. Dordrecht: Reidel, 81–108.

CHAKRAVARTTY, A. (2005). 'Causal Realism: Events and Processes', *Erkenntnis* 63: 7–31.

CHOI, S. (2002). 'Causation and Gerrymandered World Lines: A Critique of Salmon', *Philosophy of Science* 69: 105–17.

—— (2003). 'The Conserved Quantity Theory of Causation and Closed Systems', *Philosophy of Science* 70: 510–30.

COLLIER, J. (1999). 'Causation is the Transfer of Information', in H. Sankey (ed.), *Causation and Laws of Nature*. Dordrecht: Kluwer, 215–45.

DOWE, P. (1992). 'Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory', *Philosophy of Science* 59: 195–216.

—— (1995). 'Causality and Conserved Quantities: A Reply to Salmon', *Philosophy of Science* 62: 321–33.

—— (1999). 'Good Connections: Causation and Causal Processes', in H. Sankey (ed.), *Causation and Laws of Nature*. Dordrecht: Kluwer, 247–63.

—— (2000). *Physical Causation*. New York: Cambridge University Press.

—— (2001). 'A Counterfactual Theory of Prevention and "Causation" by Omission', *Australasian Journal of Philosophy* 79: 216–26.

—— (2004). 'Causes are Physically Connected to their Effects: Why Preventers and Omissions are not Causes', in C. Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell, 189–96.

EARMAN, J. (1976). 'Causation: A Matter of Life and Death', *Journal of Philosophy* 73: 5–25.

EHRING, D. (1997). *Causation and Persistence*. Oxford: Oxford University Press.

—— (2003). 'Physical Causation'. *Mind* 112: 529–33.

—— (2004). 'Counterfactual Theories, Preemption and Persistence', in P. Dowe and P. Noordhof (eds.)., *Cause and Chance*. London: Routledge, 58–76.

FAIR, D. (1979). 'Causation and the Flow of Energy', *Erkenntnis* 14: 219–50.

GARCIA-ENCINAS, M. (2004). 'Transference, or Identity Theories of Causation?', *Theoria* 19: 31–47.

HALL, N. (2004). 'Two Concepts of Causation', in J. Collins, N. Hall, and L. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 225–76.

HAUSMAN, D. (1998). *Causal Asymmetries*. New York: Cambridge University Press.

—— (2002). 'Physical Causation', *Studies in History and Philosophy of Modern Physics* 33B: 717–24.

HEATHCOTE, A. (1989). 'A Theory of Causality: Causality = Interaction (as Defined by a Suitable Quantum Field Theory)', *Erkenntnis* 31: 77–108.

HITCHCOCK, C. (1995). 'Salmon on Explanatory Relevance'. *Philosophy of Science* 62: 304–20.

—— (2001). 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98/6: 273–99.

—— (2004). 'Causal Processes and Interactions: What Are They and What Are They Good For?' *Philosophy of Science* 71: 932–41.

HUNT, I. (2005). 'Omissions and Preventions as Cases of Genuine Causation', *Philosophical Papers* 34: 209–33.

KIM, S. (2001). 'Physical Process Theories and Token-Probabilistic Causation', *Erkenntnis* 54: 235–45.

KISTLER, M. (1998). 'Reducing Causality to Transmission', *Erkenntnis* 48: 1–24.

—— (2006). *Causation and Laws of Nature*. London: Routledge.

KITCHER, P. (1989). 'Explanatory Unification and the Causal Structure of the World', in P. Kitcher and W. Salmon (eds.), *Minnesota Studies in the Philosophy of Science*. Minneapolis: University of Minnesota Press, xiii. 410–505.

KOONS, R. (2003). 'Physical Causation', *Philosophy and Phenomenological Research* 67: 244–8.

KRAJEWSKI, W. (1997). 'Energetic, Informational, and Triggering Causes', *Erkenntnis* 47: 193–202.

LEWIS, D. (1986). *Philosophical Papers II*. New York: Oxford University Press.

—— (2004). 'Void and Object', in J. Collins, N. Hall, and L. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 277–90.

MCDANIEL, K. (2002). 'Physical Causation', *Erkenntnis* 56: 258–63.

MACHAMER, P., DARDEN, L., and CRAVER, C. (2000). 'Thinking About Mechanisms', *Philosophy of Science* 67: 1–15.

MENZIES, P. (1989). 'Probabilistic Causation and Causal Processes: A Critique of Lewis', *Philosophy of Science* 56: 642–63.

MIGUEL, H., and PARUELO, J. (2002). 'Overlapping Causal Interactions in Phil Dowe's Theory', *Analisis Filosofico* 22: 69–84.

NORTON, J. (2007). 'Causation as Folk Science', in H. Price and R. Corry (eds.),

Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited. Oxford: Clarendon.

PERSSON, J. (2002). 'Cause, Effect, and Fake Causation', *Synthese* 131: 129–43.

PSILLOS, S. (2002). *Causation and Explanation*. Chesham: Acumen.

QUINE, W. (1973). *The Roots of Reference*. La Salle, Ill.: Open Court.

RIEBER, S. (2002). 'Causation as Property Acquisition', *Philosophical Studies* 109: 53–74.

REICHENBACH, H. (1958). *The Philosophy of Space and Time*. New York: Dover.

REUGER, A. (1998). 'Local Theories of Causation and the A Posteriori Identification of the Causal Relation', *Erkenntnis* 48: 25–38.

RUSSELL, B. ([1913] 1957). 'On the Notion of Cause', *Proceedings of the Aristotelian Society* 13: 1–26, repr. in *Mysticism and Logic*. Garden City, NY: Doubleday, 17–201.

—— (1948). *Human Knowledge*. New York: Simon & Schuster.

SALMON, W. (1978). 'Why Ask, "Why?"?', *Proceedings of the American Philosophical Association* 51: 683–705.

—— (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

—— (1994). 'Causality without Counterfactuals', *Philosophy of Science* 61: 297–312.

—— (1997). 'Causality and Explanation: A Reply to Two Critiques', *Philosophy of Science* 64: 461–77.

—— (1998). *Causality and Explanation*. New York: Oxford University Press.

SCHAFFER, J. (2000). 'Causation by Disconnection', *Philosophy of Science* 67: 285–300.

—— (2001). 'Physical Causation', *British Journal for the Philosophy of Science* 52: 809–13.

—— (2004). 'Causes Need Not be Physically Connected to their Effects', in C. Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell, 197–216.

SKYRMS, B. (1980). *Causal Necessity*. New Haven, Conn.: Yale University Press.

SOBER, E. (1987). 'Explanation and Causation', *British Journal for the Philosophy of Science* 38: 243–57.

—— (1988). 'The Principle of the Common Cause', in J. Fetzer (ed.), *Probability and Causality: Essays in Honor of Wesley C. Salmon*. Dordrecht: Reidel, 211–29.

THALOS, M. (2002). 'The Reduction of Causal Processes', *Synthese* 131: 99–128.

TWARDY, C. (2001). 'Physical Causation', *Philosophy of Science* 68: 266–8.

VENN, J. (1866). *The Logic of Chance*. London: Macmillan.

WOODWARD, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

CHAPTER 11
AGENCY AND INTERVENTIONIST THEORIES

JAMES F. WOODWARD

1. INTRODUCTION

Agency and interventionist theories of causation take as their point of departure a common-sense idea about the connection between causation and manipulation: causal relationships are relationships that are potentially exploitable for purposes of manipulation and control. Very roughly, if C causes E then if C were to be manipulated in the right way, there would be an associated change in E. Conversely, if there would be a change in E, were the right sort of manipulation of C to occur, then C causes E. Accounts of causation in this vein have been defended by Colling-wood (1940), Gasking (1955), von Wright (1971), Menzies and Price (1993), and Woodward (2003), among others. Similar ideas are defended by many social scientists and by some statisticians and theorists of experimental design. For example, in their influential text, *Quasi-Experimentation*, Cook and Campbell (1979: 36) 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' (emphasis in original).

Writers who have developed computational models of causal inference within a Bayes net framework, including Judea Pearl (2000) and Peter Spirtes, Clark Glymour, and Richard Scheines ([1993] 2000), have also stressed the connection between causation and manipulation, although their interest is more in the discovery of causal relationships and the prediction of the effects of manipulations than in appealing to the notion of manipulation to provide a general account of causation.

A manipulationist approach to causation is appealing in part because it appears to provide a natural treatment of the difference between causal and purely correlational claims and why we should care about this difference. As Cartwright (1983) observes, there is a correlation between, on the one hand, whether (P) one purchases life insurance from TIAA-CREF (which furnishes life insurance to college teachers) or from some other commercial life insurance company, and, on the other hand, longevity (L): purchasers of TIAA-CREF insurance tend to live longer. However, it does not follow from this observation (and it is presumably false) that purchasing TIAA-CREF insurance is a way of manipulating longevity, or that such a purchase is a means or 'effective strategy' for increasing lifespan. A manipulationist account identifies the question of whether P causes L with the question of whether some appropriate manipulation of P will be associated with a change in L; we care about whether (1) P causes L or (2) P is merely correlated with L at least in part because (1) has very different implications from (2) for whether we can manipulate L by manipulating P. This is presumably one reason

why approaches that stress the connection between manipulation and causation have been popular within experimentally oriented disciplines such as psychology and molecular biology and within disciplines that provide policy recommendations, such as economics.

Despite these appealing features, recent philosophical discussion has been largely unsympathetic to manipulationist theories. In particular, critics have claimed both that such theories are *circular* in an unilluminating way and that they are unduly *anthropocentric*. (See e.g. Hausman 1986; 1998). The charge of circularity arises because, on the face of things, the notion of manipulation looks like a causal notion; to manipulate something is to *cause* it to be in some state. How then can we appeal to the notion of manipulation to elucidate the notion of causation? The charge of anthropocentrism arises because at least on many common understandings of 'manipulation' this notion is tied to activities that human beings can carry out. There is thus a prima facie problem in extending a manipulation-based account to examples involving causal relationships in which there is no possibility of manipulation by human beings. For example, what sense can a manipulationist account give to causal claims about the effects of gravitational attraction between galaxies (in producing clumping and other large-scale structures), given that the manipulation of their causally relevant features (masses and distance from one another) is unlikely ever to be possible for humans?

It is useful to divide manipulation-based accounts into two broad categories. *Agency* theories stress the connection between causation and distinctively human agency (that is, actions and manipulations of a sort that might be carried out by human beings). Some defenders of agency theories (e.g. von Wright, Menzies, and Price) claim that one of the attractions of such theories is that they provide a way of avoiding the charge of circularity. Their idea is that the concept or experience of human agency gives us independent access to (or purchase on) the notion of causation, because the notion of agency is either not a causal notion at all or at least does not presuppose all the features of a full-blooded notion of causation. However, as we shall see, reliance on a non-causal notion of agency does not seem to yield a normatively acceptable account of causation. Because of this, several more recent accounts (e.g. Pearl 2000; Woodward 2003) that focus on the connection between causation and manipulation have dropped any appeal to distinctively human agency and instead focus on a more abstract notion of manipulation, often called an *intervention.*

2. BACKGROUND

In the remarks that follow, I focus first on agency theories and then on intervention-based (hereafter interventionist) theories. Before doing so, however, some additional stage setting is in order. I have been speaking very loosely of a general connection between causal claims and claims about what will happen under manipulations. Obviously, there are many different sorts of causal claims: there are claims that one general type or kind of factor causes another ('impacts of rocks cause bottles to break'), so-called token causal claims that the occurrence of some particular event caused another ('the impact of the rock thrown by Suzy at 3 p.m. on 9 Jan. 2005 caused this particular bottle to shatter') and so on. Within a broadly manipulationist framework, we should expect that different sorts of causal claims will be connected in different ways to claims about the outcomes of possible manipulations. Section 7 provides illustrations, but in earlier sections I will abstract away from the differences among different

sorts of causal claims when these do not seem relevant to the points I wish to make.

A second point concerns the general form that a manipulationist account should take. It seems uncontroversial that the claim that C causes E can be true even if C is not actually manipulated—any account that suggests otherwise is a non-starter. This observation suggests that manipulationist accounts should be formulated as *counterfactual* claims connecting causal claims to claims about what would happen *if* certain manipulations *were* to be performed. This is what I did in sect. 1 above. On this construal, when we infer to a causal conclusion on the basis of evidence deriving just from passive observation (that is, the evidence is not generated by experimentation) we are attempting to infer what would happen were the appropriate experiment to be performed.

A third and deeper point concerns the notion of manipulation itself. Suppose that both lung cancer L and nicotine-stained fingers N are caused by smoking cigarettes S, but L does not cause N or vice-versa. In other words, the causal relationships are such that they can be represented by the structure at Fig. 11.1, commonly called a directed graph, in which an arrow directed from X into Y means that X 'directly' causes Y and the absence of such an arrow means that X does not directly cause Y (for some additional brief remarks about directed graphs, see sect. 5).
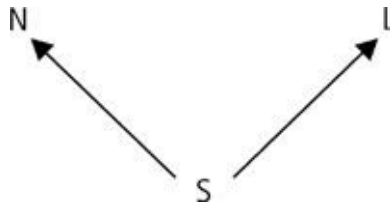
Fig. 11.1

Suppose that we decide to manipulate N by manipulating S (e.g. we force or otherwise induce some subjects to smoke and prevent others from doing so). N and L will be correlated under this manipulation of N. Nonetheless, by hypothesis, N does not cause L, in apparent contradiction to the connection between causation and manipulation on which manipulability theories rest.

This example shows that if the manipulationist approach is to be even remotely plausible, restrictions need to be imposed on what counts as an acceptable manipulation for purposes of the theory. In fact, such restrictions have a natural motivation in the theory of experimental design: everyone agrees that the manipulation just described is a badly designed experiment for the purposes of determining whether N causes L. Intuitively what we need to do to determine whether N causes L is (among other things) to manipulate N in a way that is suitably independent of (some relevant subset of) other possible causes of L such as S. More generally, we want our manipulation of N to be of such a character that if any change in L occurs, it can only occur because of the manipulation of N and not in any other way. One way of achieving this would be by means of a randomized experiment. Suppose that we have a population of both smokers and non-smokers and that depending on the output of some random device, we either assign subjects to a treatment group in which their fingers are caused to be nicotine

stained or a control group in which fingers are not so stained. Because the assignment of *N* is based on the randomizing device, it will be causally and statistically independent of *S*. We would expect that under this sort of manipulation of *N*, *N* and *L* will no longer be correlated, indicating that *N* does not cause *L*.

Manipulations of a target variable *X* that have the right sort of special characteristics to figure in a well-designed experimental test of whether *X* causes some second variable *Y* are called *interventions* in the recent literature (cf. Spirtes, Glymour, and Scheines [1993] 2000; Pearl 2000; Woodward 2003). We will explore below some alternative proposals for how best to characterize this notion but it should be apparent even at this point that understanding this notion will be central to the development of a plausible version of a manipulationist theory. One of the issues on which we will focus below is whether we can formulate a normatively acceptable notion of intervention just in terms of agency-related notions.

3. AGENCY THEORIES

By far the most detailed recent statement of an agency theory is due to Menzies and Price (1993; see also Price 1991). They propose that:

> An event *A* is a cause of a distinct event *B* just in case bringing about the occurrence of *A* is an effective means by which a free agent could bring about the occurrence of *B*.

This connection between agency and causation is used to motivate a version of a probabilistic theory of causation according to which '*A* causes *B*' is identified with '*A* raises the probability of *B*', where the probability in question is an 'agent probability':

> agent probabilities are to be thought of as conditional probabilities, assessed from the agent's perspective under the supposition that antecedent condition is realized *ab initio*, as a free act of the agent concerned. Thus the agent probability that one should ascribe to *B* conditional on *A* ... is the probability that *B* would hold were one to choose to realize *A*. (Menzies and Price 1993: 190)

In other words, the agent probability of *B* conditional on *A* is the probability of *B* conditional on the assumption that *A* has a particular sort of causal history—that *A* is realized by a free act. We can see what Menzies and Price intend by reference to the example above: their idea is when whether a subject has nicotine-stained fingers or not (*N*) is determined by a free act, any correlation between *N* and *L* should disappear—*N* does not raise the probability of *L* and hence does not cause it. Their claim is thus in effect that free acts function as interventions.

Menzies and Price hold that by appealing to the notion of agency, they can develop a non-circular, 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. ... 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. (ibid. 194-5)

Menzies and Price 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 (ibid. 194), a problem arises concerning causes for which there is no practical possibility of human manipulation. To use their own example, what can it mean to say that 'the 1989 San Francisco earthquake was caused by friction between continental plates' (ibid. 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 these 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 is causally related even though they may not be manipulable. (ibid. 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.   (ibid.)

4. PROBLEMS WITH AGENCY THEORIES

I will suggest below that these remarks embody an important psychological insight about the role of the subject's own action in causal learning. However, as it stands, the theory doesn't seem to yield a normatively correct treatment of the causal judgements we are justified in making. The crux of the difficulty is that while there is indeed reason to think that the extensions of the concepts 'event due to a free action' and 'event satisfying the conditions for an intervention' overlap to some significant extent (cf. sect. 11), these concepts are very far from being exactly coextensive—an action may be free, at least in the senses normally recognized by philosophers, and yet fail to satisfy the conditions for an intervention. If, in such cases, we take correlations between X and Y that persist under free acts that realize X to show that X causes Y, we will often reach mistaken causal conclusions. Conversely, an action may not be free (or some process may occur which is not a human action at all) and yet the process may still satisfy the conditions for an intervention. If, when X is produced by such a process, X and Y remain correlated, we may conclude that X causes Y.

To illustrate, suppose that a human experimenter is faced with a common cause structure such as that of Fig. 11.1 and 'freely chooses' to manipulate N in a way that is correlated with the common cause S. (Perhaps the experimenter does this by observing whether individual subjects smoke and then manipulates N accordingly.) Note that there is nothing in the concept of a 'free action' as ordinarily understood, according to which the experimenter's actions are 'unfree' simply because they are influenced by or correlated with S. Then the occurrence of nicotine-stained fingers, when produced by such free actions, will raise the probability of lung cancer, even though the former does not cause the latter. Could we respond to this difficulty by making it part of the characterization of a 'free action' A that produces N that A is not correlated with other causes of N? A moment's thought shows that this additional condition is far too strong. As long as the free action itself has some causes (as it will on a non-libertarian account of free will) or as long as there are causally intermediate events between the free action A and N, these will be 'other causes' of N that are correlated with A. Moreover, the proposed condition is inadequate in other respects as well: for example, if the free act A itself directly causes both N and the other joint effect S (via a route that does not go through N), again the condition will not rule out the mistaken conclusion that N causes L.

More generally, it seems obvious that an adequate statement of the condition that Menzies and Price are looking for will need to make reference not just to N but to the putative effect L of N—as noted above, we need to exclude the possibility that this putative effect is the result of some causal pathway that does not go though N. It is unclear how the agency-related notions invoked by Menzies and Price might be used to make the distinctions needed to do this —instead, the needed distinctions appear to be overtly causal in character, and this undermines the reductionist aspect of Menzies and Price's project.

I turn next to a second set of problems for agency theories, having to do with the fact that our (i.e. human) notion of causation is such that we readily think of causal relationships as holding in contexts in which we have no 'personal experience' of agency. As the passages quoted above make clear, Menzies and Price think of this as primarily raising a problem having to do with causes that humans are unable to manipulate. In fact, however, there is a prior and in some respects more fundamental problem that arises even for readily manipulable causes. Consider the contrast between two scenarios. In the first, you throw a rock which

strikes a second rock, setting it in motion. Here you presumably have direct personal experience of your agency in setting the first rock in motion, but no such experience with respect to the second rock. In the second scenario, you observe one rock, set in motion by some natural cause such as the wind, strike a second rock with the result that it moves. Our concept of cause is such that we think that in both cases the impact of the first rock causes the second rock to move, but why exactly is it, on Menzies and Price's theory, that we are entitled to this judgement? After all, in the second scenario, I presumably have no experience of my own agency at all and even in the first, my experience of agency seems limited to producing the motion of the first rock. Within an agency theory, what justifies grouping the relationship between the movement of my hand and the motion of the first rock in the first scenario and the relationship between the movements of the first and second rocks under the common rubric of 'causation'?

It might seem that one possible answer (call this the 'projection hypothesis) is that in thinking of the movement of the first rock as a cause of the movement of the second, subjects consciously or unconsciously transfer their own experience of agency to the first rock—that is, they think of the first rock as (in some way) an agent which 'acts' on the second rock.[1] However, as it stands, this suggestion is not very satisfying. To begin with, we need more details about how this projection process works and why people engage in it, especially since the attributions in question, if taken literally, are so obviously mistaken—rocks are not really agents and so on. If the idea is that agents somehow find it useful to think about causal interactions involving rocks 'as if' they involved agents, even though such reasoning is not literally correct, we need some story about *why* this is useful and how this (mistaken) reasoning allows us capture the contrast between (what we normally think of) as true and false causal claims involving inanimate objects.[2] We also need to ask how the projection hypothesis might be tested and what evidence supports it. On one natural construal, the hypothesis predicts that the brain areas/psychological processes involved in the agent's own sense of agency and attribution of mental states to others are also centrally involved when agents attribute causal influence to inanimate objects. This prediction appears to be false.[3]

Notice that this is *not* an issue about unmanipulable causes—it may be easy for me to manipulate the rocks in the second scenario if I choose to do so. The problem is rather that within Menzies and Price's framework we need some empirically grounded account of the processes of inference, analogical reasoning, imaginative extension, and so on (the processes that underlie projection) that lead from the fundamentally first-person experience of agency to a concept of cause that does not seem to require this experience for its application. It is interesting to note (cf. sect. 11) in this connection that several authors have suggested that many non-human animals, including other primates, have a grasp of an egocentric cause-like notion in the sense that they are capable of learning relationships between their own actions and the outcomes those actions produce, but that they fail to grasp that the very same relationships can hold between objects and events in the world, independently of their (or anyone's) actions or experience of agency, and that this has important consequences for the causal learning and understanding they are capable of. If so, such animals may possess (or at least behave as though they are guided by) an agency-based cause-like notion resembling the notion described by Menzies and Price, but not the concept of cause possessed by adult humans.

Quite apart from the projection problem just described there is also, as Menzies and Price recognize, a distinct problem having to do with the extension of their theory to causes for which there is no possibility of human manipulation. Menzies and Price attempt to resolve this problem by appealing to the idea that cases involving non-manipulable causes 'resemble' or can be modelled by cases involving manipulable causes and that, in virtue of this resemblance relationship, we can use our grasp of the latter to understand the former. It is of course crucial to this strategy that (as Menzies and Price claim) the resemblance in question be specified in non-causal terms. If, in specifying what it means for the movements of the continental plates to 'resemble' the artificial models that the seismologists are able to manipulate, we had to appeal to *causal* similarities between the two structures, we would no longer be explaining the content of claims about unmanipulable causes in terms of claims about manipulable causes. Instead, we would be relying on an unexplained notion of causal similarity between manipulable causes understood on the basis of agency and unmanipulable causes that must be understood in some other way. However, Menzies and Price provide no reason to believe that the needed resemblance relation can be specified non-causally and there is good reason to be sceptical of this claim. 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.

5. INTERVENTIONS

Our discussion so far has shown that if we wish to formulate a satisfactory statement of the connection between causal claims and the outcomes of ideal manipulations ('interventions'), we need to be precise about what constitutes an intervention. There have been a number of attempts to do this in the recent literature, including Spirtes, Glymour, and Scheines (1993; 2000), Hausman (1998), Pearl (2000), Woodward and Hitchcock (2003), Woodward (2003). All these writers focus on what is broadly the same idea but offer formulations that differ somewhat in detail, in part because they are animated by somewhat different theoretical purposes. I begin with Pearl (2000) who provides one of the most detailed recent attempts to think systematically about interventions and their significance for understanding causation.[4] Pearl follows a standard tradition in the econometrics and the causal modelling literature of using systems of equations to represent causal relationships. He also employs directed graphs for the same purpose. His work provides a striking illustration of the heuristic usefulness of a manipulationist framework in giving a causal interpretation for such representations. Since the notion of an intervention in Pearl's work is characterized in terms of equations and graphs, I begin with some brief remarks about these. For Pearl, a functional causal model is a system of equations $X_i = F(Pa_i, U_i)$ where X_i, Pa_i, and U_i are all variables (See sect. 6 below). Pa_i represents the direct causes or, as they are sometimes called, the 'parents' of X_i that are explicitly included in the model and U_i represents an error variable that summarizes the combined impact of all other variables that are causes of X_i. Pearl takes each equation to represent a distinct 'causal mechanism' which is understood to be 'autonomous' in the sense

in which that notion is used in econometrics; this means roughly that it is possible to interfere with or disrupt each mechanism (and the corresponding equation) without disrupting any of the others. At least for the purposes of defining the notion of an intervention the notion of a causal mechanism or direct cause is taken as primitive and the notion of an intervention is defined in terms of it.

The simplest sort of intervention in which some variable X_i is set to some particular value x_i amounts, in Pearl's words (ibid. 70), 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 keeping all other mechanisms undisturbed' (I have altered the notation slightly). In other words, the intervention disrupts completely the relationship between X_i and its parents so that the value of X_i is determined entirely by the intervention. Furthermore, the intervention is 'surgical' in the sense that no other causal relationships in the system are changed. (This is sometimes described as the 'equation wipe out' conception of interventions.) Formally, this amounts to replacing the equation governing X_i with a new equation $X_i = x_i$, substituting for this new value of X_i in all the equations in which X_i occurs but leaving the other equations unaltered. It is assumed that the other variables in the system that change in value under this intervention will do so only if they are effects of X_i.

As an illustration, consider again the common cause structure from sect. 2, which may be represented by the equations:

(5:1) $N = F_1(S, U_1)$

(5:2) $L = F_2(S, U_2)$

The effect of an intervention on the variable N is represented by replacing equation (5.1) with a new equation (5.3) $N = n_i$, indicating that N has been set by the intervention to the value n_i and is no longer causally influenced by the variable S that was previously its direct cause. The other equation in the system is undisturbed by this alteration, so that the structure of the new system in which the intervention has occurred is represented by (5.3) and (5.2). Within a framework like Pearl's, causal relationships may also be represented by directed graphs. An arrow from a variable X_i *into* a second variable X_j (that is, an arrow with X_i at the tail and X_j at the head) means that X_i is a direct cause of X_j. Thus, as we have already noted, the common cause structure (5.1)—(5.2) can also be represented by Fig. 11.1.

Interventions also have a simple graphical representation: an intervention I on variable X_j 'breaks' or removes all other arrows directed into X_j and replaces these with a structure in which the only arrow into X_j is an arrow from I. All other arrows in the graph are left undisturbed. Thus the effect of an intervention on N in Fig.11.1 is to replace this structure with that shown in Fig. 11.2, again representing that N is entirely under the control of the intervention variable and that other causal influences on N have been broken.

Pearl's talk of 'lifting' the variable intervened on from the influence of its (previous) direct causes may seem puzzling to philosophers who are accustomed to associate causal

relationships with the instantiation of (presumably inviolable) 'laws of nature' but in fact the underlying idea is quite intuitive.[5] An intervention replaces a situation in which the variable X intervened on is sensitive to changes in the values of certain variables with a new situation in which the value of X is no longer sensitive to such changes but instead depends only on the value assigned to it by the intervention.[6] Many real-life experiments aim at (and succeed) in accomplishing this. For example, in an experiment to test the impact of a drug on recovery in which subjects are randomly assigned to a treatment group that receives the drug and a control group from which the drug is withheld, the ideal at which one aims is that who receives the drug should be determined entirely by the random assignment (the intervention), and not, as it presumably was previously, by such other factors as the subject's own decisions. As a matter of methodology, the reasons for employing experiments with this feature (when this is possible) is straightforward: by severing the relationship between the variable intervened on and its previous (or 'endogenous') direct causes, we eliminate some (although not all) possible sources of 'confounding'—for example, we ensure that those previous causes are not common causes of both the variable intervened on and the putative effect. We also ensure that the variable intervened on has whatever value the experimenter intends to give it, since this value is not affected by anything other than the intervention process.
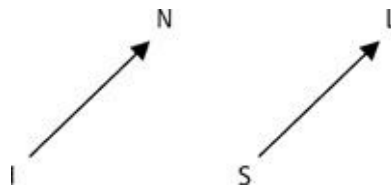
Fig. 11.2

Pearl represents the proposition that the value of X has been set by an intervention to some particular value, x, by means of a 'do' operator $do\ (X = x)$ or $do\ (x)$. This allows for simple 'definitions' (as Pearl calls them) of various causal notions. For example, the 'causal effect' of X on Y associated with the 'realization' of a particular value x of X is defined as $P(y/do\ x)$—this represents the 'total effect' of $X = x$ on Y through all different paths from X to Y. By contrast, the 'direct effect' of $X = x$ on Y is $P(y/do\ x, do\ S_{xy})$ where S_{xy} is the set of all endogenous variables except X and Y in the system. That is, the direct effect is the distribution that Y takes under an intervention that sets $X = x$ and fixes by interventions the values of all other variables in the system—according to Pearl (ibid. 126), this represents the sensitivity of Y to changes in X alone.

We can further clarify the notion of an intervention by contrasting it with the more familiar notion of 'conditioning' (on a passively observed value of a variable) (cf. Meek and Glymour 1994). In the structure from sect. 2 in which S is a common cause of L and N, L and N are unconditionally dependent. That is, the probability of L conditional on N is different from the unconditional probability of L: $P(L/N) \neq P(L)$. However, under an intervention on N, which we may represent by conditioning on $do\ N$, L and $do\ N$ will be independent: $P(L/do\ N) = P(L)$. Conditioning on an observed value of N is thus a fundamentally different operation from intervening on N. This is because[7] when we condition on the observed values of N, we assume

that whatever causal structure generates those values remains intact, while intervening on N alters the causal structure of this system. Thus if we observe the values of N and know that the value of L is generated by [Fig. 11.1](#), this provides information about L; not so if the value of N is set by an intervention. Causal claims have to do with what will happen under interventions, although, given plausible assumptions, they also will have certain systematic connections to conditioning relationships.[7]

Pearl's characterization of the notion of an intervention seems ideally suited for the purposes for which he uses it. Basically these purposes are calculational or predictive rather than the more foundational ones that motivate philosophical accounts of causation. In particular, much of the focus of Pearl's discussion is on showing how to calculate the quantitative value of (to 'identify') causal effects and to predict the effects of interventions when we have qualitative information about causal structure and information about the probability distribution of the variables in the system of interest.

Arguably, however, Pearl's characterization is less well suited to the task of using the notion of an intervention to characterize what it is for a relationship to be causal. One reason[8] for thinking this derives from Pearl's requirement that an intervention on X_i leave intact all other mechanisms besides the mechanism that previously determined the value of X_i. If, as Pearl apparently intends, we understand this to include the requirement that an intervention on X_i must leave intact the causal mechanism if any, that connects X_i to its possible effects Y, then an obvious worry about circularity arises, if we want to use the notion of an intervention to characterize what it is for X_i to cause Y. In part for this reason, Woodward and Hitchcock (2003) and Woodward (2003) explore a different way of characterizing the notion of an intervention that does not make reference to the relationship between the variable intervened on and its effects. For Woodward and Hitchcock (hereafter WH), in contrast to Pearl, an intervention I on a variable X is always defined with respect to a second variable Y (the intent being to use the notion of an intervention on X with respect to Y to characterize what it is for X to cause Y). An intervention I must meet the following requirements to count as an (WH) intervention:

(M1) I must be the only cause of X—that is, as with Pearl, the intervention must completely disrupt the causal relationship between X and its previous causes so that the value of X is set entirely by I,

(M2) I must not directly cause Y via a route that does not go through X.

(M3) I should not itself be caused by any cause that affects Y via a route that does not go through X,

(M4) I must be probabilistically independent of any cause of Y that does not lie on the causal route connecting X to Y.

Before considering how the WH notion of an intervention might be used to characterize various causal notions, let us note how both it and Pearl's notion differ from the agency-related notions to which Menzies and Price appeal. Neither Pearl's nor WH's notion involves

human agency or activity—instead both define interventions in terms of causal and (in the case of WH) correlational relationships. A purely natural process, not involving human activity at any point, will count as an intervention as long as it has the right causal and correlational characteristics. This allows such intervention-based accounts to avoid at least some versions of the charge of anthropomorphism, although as we shall see (sect. 10) there still remain questions about their range of application. However, since the characterization of an intervention is overtly causal, it may seem that worries about 'circularity' in interventionist accounts become even more pressing—at the very least it is clear that we cannot appeal to Pearl's or WH's notion of an intervention to give a reductive account of what it is for a relationship to be causal. I will address this worry about circularity below, but I want first to explore in more detail how the notion of an intervention can be used in the characterization of causal relationships.

6. CAUSATION AND INTERVENTIONS

Within a manipulationist framework, causes and effects must be manipulable, at least 'in principle'. This in turn suggests that we should think of causal relata as capable of varying or of being in a range of different possible states or conditions. It is thus natural to think of causal claims as having to do with relationships between *variables*, where the mark of variable is that it is capable of taking more than one value. I have already employed this convention in connection with many of the examples discussed above and the use of variables in the representation of causal relationships is standard practice in many areas of science. Our initial focus will be on type causation and on capturing a broad notion of causal relevance that corresponds to the idea of one factor being positively, negatively, or of mixed causal significance for another. The usual assumption in the philosophical literature that causation is a relationship between events or event types can be readily captured within this variable-based framework in terms of 'indicator' or two valued variables corresponding to the occurrence or non-occurrence of the events of interest. Thus we may express the causal claim that short circuits cause fires in terms of a relationship between two variables S and F, with S taking two possible values corresponding to the occurrence or non-occurrence of a short circuit, and F taking two possible values corresponding to the occurrence or non-occurrence of the fire.

Consider now the following proposals that give candidates for necessary and sufficient conditions for 'X causes Y' where X and Y are variables and 'causes' means (as explained above) 'is causally relevant to':

(SC) If (i) there are possible interventions that change the value of X such that (ii) under such interventions (and no others) X and Y are correlated, then X causes Y.

(NC) If X causes Y then (i) there are possible interventions that change the value of X such that (ii) under such interventions (and no other interventions) X and Y are correlated.

The causal notion captured by NC and SC is relatively weak and uninformative. It

corresponds to the question 'is *X* causally relevant to *Y* at all', where this is interpreted as the question of whether there is there *some* change in the value of *X* which will change the value of *Y* or the probability distribution of *Y*. We are, of course, also interested in the elucidation of more precise causal claims having to do with the exact way in which *X* is causally relevant to *Y*—which from a manipulationist perspective has to do with exactly which changes in *X* will be associated with which changes in *Y* and under what conditions. As we shall see, the content of such claims may be captured within a manipulationist framework by extending the characterizations below in obvious ways.

SC 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 that are sometimes regarded as necessary for causation. This is a highly non-trivial claim. It implies, for example, that 'double prevention' (Hall 2000) or 'causation by disconnection' (Schaffer 2000) involves genuine causal relationships (because these are relationships that support manipulation) even though the cause is not connected to its effect via a spatio-temporally continuous process and even though there is no transfer of energy and momentum from cause to effect. Similarly, if an 'action at a distance' version of Newtonian gravitational theory had turned out to be correct, this would be a theory that described genuine causal relationships, on an interventionist account of causation. This illustrates one respect in which an interventionist theory will reach very different conclusions about which relationships are causal from other competing theories.

What about NC? Consider the causal structure represented by means of the equations

(6:1) $Y = aX + cZ$

(6:2) $Z = bX$

and by the associated directed graph shown in Fig. 11.3.

If $a = -bc$, the direct causal influence of *X* on *Y* will be exactly cancelled out by the indirect influence of *X* on *Y* that is mediated through *Z*. If it is nonetheless correct to think that *X* (in some relevant sense) causes *Y,* then NC will be false, since there are no interventions on *X* alone that will change *Y*.[9]

This example shows that we need to distinguish between two notions of 'cause'.[10] 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 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 or directed path even if, because of cancellation, *X* has no total effect on *Y*. While both SC and NC are plausible if 'cause' is interpreted as 'total cause' (where, it should be recalled the causal notion we are trying to capture is a broad notion of a causal relevance), NC is not plausible if 'cause' is interpreted as 'contributing cause' although SC remains plausible under this interpretation.
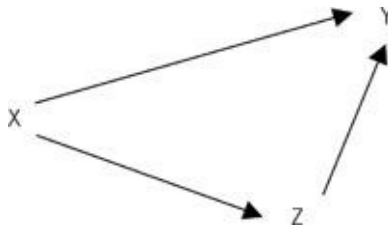
Fig. 11.3

Can we capture the notion of a contributing cause within a interventionist framework? The strategy followed in Woodward (2003) is to first formulate 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. Woodward characterizes direct causation thus:

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

Note that this characterization appeals to what will happen to Y under *combinations* of interventions, on both X and on other variables besides X. For example, X is a direct cause of Y in Figure 11.3 because, if we intervene to fix the value of Z and then intervene to change the value of X in a way that is independent of the intervention on Z, the value of Y will change. This contrasts with the characterization of total cause which appeals just to what will happen to Y under an intervention on X alone and no other variables. Note also that at this point we have moved well beyond earlier formulations of agency and manipulability theories that attempt to characterize causal relationships by appealing just to what will happen under *single* interventions on the cause variable.

Using DC we may formulate a necessary and sufficient condition, expressed in terms of claims about the outcomes of hypothetical interventions, for X to be a contributing, (type-level) cause of Y:

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

(Motivation for this definition as well as illustrative examples are given in Woodward 2003.)

7. OTHER CAUSAL NOTIONS

We noted above that the causal claims characterized by NC, SC, and M are in one sense very weak—they refer only to there being some correlation between X and Y under some interventions on X. There are a variety of ways in both science and common sense that more detailed and specific causal information may be conveyed and these also have a natural interpretation within an interventionist framework. These include the formulation of quantitative relationships, represented by functions. For example, the relationship between the extension X and restoring force F exerted by a particular type of spring $F = -kX$ tells us exactly how for some range of interventions that alter X, F will change. Various qualitative locutions may be used for a similar purpose. For example, we often use 'causes' to express the idea that one factor X is a positive or promoting causal factor for another factor Y (rather than just being causally relevant to Y), as when we say that smoking causes lung cancer. Depending on the details of the case, such locutions may be interpreted within an interventionist framework as claims about various qualitative features of the functional form linking cause and effect—for example, that for variables representing smoking S and lung cancer L that take two possible values {present, absent}, an intervention that changes the value of S from absent to present increases the probability that the value of L will be present rather than absent. Alternatively, if S and L are understood as more quantitative variables—for example, if S is measured by average number of cigarettes consumed per week and L by probability of lung cancer, what might be intended by the claim that S is a promoting cause of L is that L is a monotonically increasing function of S, at least over much of its domain.

In still other cases, the use of *contrastive focus* will provide a natural way of conveying information about how manipulation of the cause will alter the effect. Suppose that the presence of potassium salts in a warehouse fire causes the flames to be purple. This claim might be interpreted as meaning something like: an intervention that changes whether potassium salts are present (when a fire occurs) will be associated with a change of colour in the flames. However (barring special circumstances) such an intervention will not change whether a fire occurs (rather than not occurring)—that is (as we might say), the presence of the salts causes the fire to be purple rather than some other colour, but does not cause the fire to occur.

In general, then, as these examples illustrate, within an interventionist framework spelling out the content of detailed and specific causal claims will be a matter of specifying exactly which interventions on the cause variable will be associated with which changes in the effect variable and under which background circumstances. In this respect, an interventionist approach is more general than many other accounts in the philosophical literature that are predicated on the assumption that causal claims must assume some more specific canonical form—for example, accounts that assume that all causal claims must relate binary 'events' or must relate random variables with a well-defined joint probability distribution, as so-called probabilistic theories of causation do.

So far our focus has been on type causal claims. There are also several proposals in the literature that provide interventionist treatments of token or actual cause claims. For reasons of space, I will not attempt to describe these proposals in detail but will merely gesture at the

basic idea, which is to appeal to what will happen to the effect under combinations of interventions that both affect the cause and fix certain other variables to specific values. It is worth noting that accounts taking this form are able to deal in a reasonably intuitive fashion with many of the standard counterexamples to certain other treatments of token causation. Suppose gunman 1 shoots (s_1) victim causing his death (d), where gunman 2 does not shoot but would have shot (s_2) also causing d, if s_1 had not occurred. If we fix the behaviour of gunman 2 at its actual value (he does not shoot), then an intervention that alters whether gunman 1 shoots will alter whether victim dies, thus identifying s_1 as the actual cause of d, despite the absence of counterfactual dependence (of the usual sort) between d and s_1.[11]

Although this appeal to combinations of interventions may strike some as artificial, in fact it maps onto standard experimental procedures in a natural way. Consider a case of genetic redundancy—gene complex G_1 is involved in causing phenotypic trait P but if G_1 is inactivated another gene complex G_2 (which is inactive when G_1 is active) will become active and will cause P. The geneticist may test for this possibility by first interfering with G_2 so that it is rendered permanently inactive and then intervening to vary G_1 and observing whether there is a corresponding change in P, and, second, intervening to render G_1 inactive and then, independently of this, turning G_2 on and off and observing whether there is a change in P. As this example illustrates, we may think of different complex causal structures in which there are multiple pathways, redundancy, cancellation, and so on, as encoding different sets of claims about what will happen under various possible combinations of interventions.

8. THE PROBLEM OF CIRCULARITY

Suppose, as argued above, that plausible versions of the manipulationist approach must appeal to a notion of intervention that is itself causal in character. How damaging is this to such accounts? The answer will depend in part on what we take the legitimate goals and aspirations of an account of causation to be. Many philosophers have supposed that an acceptable theory of causation must be reductionist—that is, it should explain causal notions in terms of concepts that are not themselves causal and that meet certain agreed-upon criteria for intelligibility and testability. Typically these criteria are broadly empiricist—thus it is assumed that the reduction will involve such non-causal concepts as regularity, spatio-temporal contiguity, and the like. It is obvious that an intervention-based account of causation will not be reductive in this sense. On the other hand, there is no generally accepted reductive account of causation and a number of writers (e.g. Cartwright 1983) have argued that there are good reasons for supposing that no such account is possible. In addition, there are many examples from both science and common sense of interrelated families of concepts that do not seem reducible to concepts that lie outside such families and yet seem nonetheless to satisfy reasonable standards of intelligibility and testability—'probability' is a standard illustration. This suggests that we can often make real progress in elucidating some concept of interest by showing how it connects up with other concepts and how claims involving it can be tested even if we cannot provide a non-circular reduction. Advocates of interventionist theories can claim that a similar point holds for 'cause'—even if we cannot reduce the various versions of

this concept to something else, we can elucidate its content by showing how it connects up with other causally based concepts such as intervention, and how claims involving it can be tested both experimentally and otherwise.

It is thus worth asking those who require that an acceptable account of causation must be reductionist just what the motivation or rationale for this requirement is. If we think of various theoretical concepts (such as 'electron') that figure in scientific theories, the idea that they must be definable in terms of or reducible to some other set of supposedly more empirically legitimate concepts (e.g. 'observable concepts') was abandoned a long time ago as indefensible 'concept empiricism'. On the face of things, those who contend that any acceptable account of causation must be reductionist are urging the analogue of concept empiricism for 'causation'. They need to explain more clearly than they have hitherto why this demand is in order in the case of 'causation' even though it has been given up in other cases.[12]

A related point is that it is simply a mistake to suppose that because manipulability approaches are non-reductive, they are trivial, tautological, or lacking in interesting content. For one thing, as the discussion in sect. 6 shows, manipulability accounts can *conflict* with other accounts of causation, leading to different causal judgements in particular cases (e.g. in cases in which there is action at a distance, double prevention, etc.). In addition, the issue of how best to characterize the notion of an intervention and how to connect it to causal claims in such a way as to avoid obvious counterexamples (such as those discussed in sect. 2) is a highly nontrivial matter. Moreover, as we have noted, there are a number of different causal concepts —total causation, direct causation, token causation, and so on. Even within a broadly interventionist framework, we face a number of non-trivial choices about how such concepts connect to each other, and to the notion of intervention. An interventionist framework can thus be very useful in exhibiting the differences and interconnections among different causal concepts even if it fails to be reductive. Note also that although defenders of an interventionist account are committed to the idea that such an account can be worthwhile and illuminating without being reductionist, there is nothing in the interventionist approach per se that excludes the possibility that such an account might be supplemented or complemented by some other approach that does offer a reduction of key interventionist concepts such as 'intervention'. Interventionists may be sceptical that such an account will ever be forthcoming, but they need not reject its possibility a priori.

There is yet another observation that bears on the issue of circularity. Note that although the WH characterization of an intervention I on X with respect to Y does make use of causal information, this is *not* information about the existence or nonexistence of a causal relationship *between X and Y*. Instead the information concerns the causal relationship between I and X, between I and other causes of Y besides X, and so on. In other words, the WH characterization connects information about *other* causal relationships besides the $X \rightarrow Y$ relationship and correlational information to a claim about what must be true for X to cause Y. The characterization of an intervention on X with respect to Y is thus not viciously circular in the sense that it presupposes the very thing we are trying to elucidate—whether there is a causal relationship from X to Y. Regardless of whether the WH characterization or other characterizations found in the literature are fully adequate, there is a very compelling reason for thinking that *some* non-viciously circular characterization must be possible: we do after all

learn about causal relationships by performing relatively black-box experiments, and it is not easy to see how this is possible unless we can sometimes recognize whether there has been an intervention on X with respect to Y without presupposing an answer to the question of whether X causes Y.

9. IN WHAT SENSE MUST INTERVENTIONS BE POSSIBLE?

SC, NC, and M refer to 'possible interventions'. There is a range of ways this phrase might be interpreted, corresponding to more or less strict notions of possibility. Note first that because the notion of an intervention has been given a non-anthropomorphic characterization, there is nothing in the versions of interventionist theory formulated above that motivates restriction of 'possible interventions' to interventions that are within the present practical or technological powers of human beings. As long as we can sensibly entertain counterfactuals about what would happen to Y if some natural process meeting the conditions for an intervention were to occur on X, we can apply the interventionist theory. This will certainly include some large range of cases in which the interventions in question are of such a character that they cannot at present be carried out by humans. Matters become stickier, however, when we consider circumstances in which the relevant notion of an intervention is physically or nomologically impossible. Consider (cf. Woodward 2003) the claim that the gravitational attraction F_m of the moon causes the behaviour of the tides. It is arguable that not only is it not technologically possible for humans to change the value of F_m (e.g. by changing the position of the moon) but that any physically possible process that might accomplish this would violate the conditions for an intervention, roughly because the process would not be sufficiently 'surgical'. For example, if nature were to change the position of the moon by introducing a new gravitating body in its vicinity, this body would exert an independent gravitational influence of the tides in violation of the requirement in the characterization of an intervention. Woodward (2003) responds by suggesting that in at least some cases of this sort (including the one under discussion) we are in possession of a well-confirmed theory (Newtonian gravitational theory) that tells us what would happen in the (arguably) contra-nomic circumstances in which the moon occupies a different position as a result of an intervention and this is sufficient for the evaluation of the appropriate counterfactuals. Others may think, however, that at this point we have moved beyond the most natural range of application of the manipulationist theory. I will return to this issue in sect. 11.

Finally, consider cases in which interventions may be impossible or ill-defined for conceptual or metaphysical reasons. For example, some hold that there is no well-defined process of changing an animal of one biological species into a member of some other biological species—if so, claims like 'N being a tiger causes N to run fast' will lack an interventionist interpretation. Several prominent statisticians who favour manipulationist accounts of causation (e.g. Rubin 1986; Holland 1986) have argued on similar grounds that claims attributing causal efficacy to race and gender are not meaningful. Others (e.g. Glymour 2004) agree that such candidates for causes are unmanipulable in principle but are also antecedently convinced that causal claims involving them are meaningful and hence take such examples to reflect an important limitation on the scope of manipulationist accounts.

However, as Woodward (2003) argues, causal claims involving unmanipulable causes often *are* unclear in meaning or ambiguous and that their meaning can often be clarified by replacing them with similar but related claims involving manipulable causes. This is just what one would expect if a manipulationist account of causation is correct.

10. SCOPE OF INTERVENTIONIST ACCOUNTS

In sect. 9, we observed that although it is natural to formulate an interventionist account in terms of counterfactuals about what would happen under possible interventions, it is arguable that as we make the relevant notion of 'possible intervention' more and more permissive, so that it includes contra-nomic possibilities and so on, we reach a point at which this notion and the counterfactuals in which it figures become so unclear that we can no longer use them to illuminate or provide any independent purchase on causal claims. It is an interesting and unresolved question whether the point at which this happens is also the point at which the associated causal claims no longer strike us as clear or useful, which is what one would expect if interventionism is a complete account of causation.

This issue arises in a particularly forceful way when we attempt to apply such accounts to fundamental physical theories understood as applying to the whole universe. Consider this claim:

(10.1) The state S_t of the entire universe at time t causes the state S_{t+d} of the entire universe at time $t + d$.

where S_t and S_{t+d} are specifications in terms of some fundamental physical theory.

On an interventionist construal, (10.1) is unpacked as a claim to the effect that under some possible intervention that changes S_t, there would be an associated change in S_{t+d}. The obvious worry is that it is unclear what would be involved in such an intervention and unclear how to assess what would happen if it were to occur, given the stipulation that S_t is a specification of the entire state of the universe. How, for example, might such an intervention be realized, given that there is nothing left over in addition to S_t to realize it with?

Commenting on an example like this, Pearl (2000: 350) writes, 'If you wish to include the whole universe in the model, causality disappears because interventions disappear—the manipulator and the manipulated lose their distinction.' Whether or not Pearl is right about this, it seems uncontroversial that it is far from straightforward how to interpret the interventionist counterfactual associated with (10.1). The interventionist account seems to apply most naturally and straightforwardly to what Pearl calls 'small worlds'—cases in which the system of causal relationships in which we are interested is located in a larger environment which serves as a potential source of outside or 'exogenous' interventions. The systems of causal relationships that figure in common-sense causal reasoning and in the biological, psychological, and social sciences all have this character but fundamental physical theories do not, at least when their domain is taken to be the entire universe.

There are several possible reactions to these observations. One is that causal claims in

fundamental physics such as (10.1) are literally true and that it is an important limitation in interventionist theories that they have difficulty elucidating such claims. A second, diametrically opposed reaction, which I take to be Pearl's, is that causal concepts do not apply, at least in any straightforward way, to some or many fundamental physics contexts and that is a virtue of the interventionist account that it helps us to understand why this is so. This second suggestion may seem deeply shocking and counterintuitive to philosophers who believe that all causal claims must be 'grounded' in ('made true by') fundamental physical laws. In fact, however, the view that fundamental physics is not a hospitable context for causation and that attempts to interpret fundamental physical theories in causal terms are unmotivated, misguided, and likely to breed confusion is probably the dominant, although by no means universal, view among contemporary philosophers of physics.[13] According to some writers (Hitchcock 2007*a*; Woodward 2007*a*), we should take seriously the possibility that causal reasoning and understanding apply most naturally to small world systems of medium-sized physical objects of the sort studied in the various special sciences and look for an account of causation, such as the interventionist account, that explains this fact. The question of the scope of interventionist theories and causal claims in general is thus an important and at present unresolved issue.[14]

11. AGENCY AND INTERVENTIONIST ACCOUNTS IN PSYCHOLOGICAL PERSPECTIVE

So far our focus has been on the evaluation of agency and interventionist accounts as normative theories of causal inference and judgement. However, both theories also can be interpreted as suggesting various descriptive claims about the empirical psychology of causal learning and judgement among both humans and non-humans. For example, Menzies and Price's version of the agency theory is, inter alia, a theory about the origins of causal concepts in humans. There are very rich and rapidly growing literatures within (human) cognitive and developmental psychology, primatology, and animal learning that bear on these empirical claims. Because of space constraints, I can only gesture at a few themes within this literature.[15] First, a number of writers have noted the close similarity between instrumental or operant (as opposed to classical) conditioning and causal learning when viewed from a manipulationist perspective. In instrumental conditioning, what is learned is an association between some behaviour produced by the subject and an outcome, as when rats learn an association between pressing a lever and the provision of a food pellet. There are striking, if incomplete, parallels between instrumental conditioning in non-human animals and causal learning and judgement in humans—for example, human judgements of causal strength are subject to discounting effects when alternative causes are present, and exhibit backward blocking, just as instrumental conditioning does. In general, humans behave as though estimates of the instrumental efficacy of their action tracks causal efficacy, which is what one would expect on a manipulationist theory of causation.

A second theme concerns the role of a subject's own actions in facilitating causal learning. There is a great deal of evidence that the ability to intervene or manipulate facilitates causal learning in both adults and small children in comparison with passive observation. Both groups are able to reason to normatively correct causal conclusions in cases involving both a

single intervention and combinations of interventions, and to distinguish between intervening and conditioning in normatively appropriate ways. These observations suggest that interventionist accounts capture something that is 'psychologically real' in human causal judgement. Although, for reasons explained above, it is dubious that the human concept of cause is derived just from the experience of agency, it is a natural interpretation of the experimental evidence that this experience plays an important role in learning particular causal relationships. Woodward (2007*b*) suggests that humans including infants have (1) a default tendency to behave or reason as though they take their own voluntary actions to have the characteristics of interventions and (2) associated with this a strong tendency to take changes that temporally follow those interventions with a relatively short delay as caused by them. We see evidence for this tendency in the existence of well-known causal illusions in which we experience salient changes that follow our voluntary actions as caused by them. Of course, as noted above, by no means all voluntary actions qualify as interventions, but nonetheless it may be that people have a defeasible tendency to assume this and that this tendency facilitates causal learning, especially in young children. Thus, even if agency-based accounts do not yield a normatively adequate account of causation, they may have a great deal of value as accounts of the acquisition of causal knowledge.

A third issue concerns the relationship between the sorts of capacities/causal understanding that are manifested in the ability to intervene and manipulate and other capacities that are often associated with causal understanding or possession of a concept of causation. For example, a number of psychologists (e.g. Leslie and Keeble 1987) and philosophers (Prinz 2002) claim that the visual responses (as measured by looking time) of infants to so-called launching phenomena (mechanical collisions of the sort studied by Michotte involving the perception of causation) and to object permanence tasks show that even very young children possess a concept of causation and are capable of causal reasoning. Although the evidence is controversial, it is widely believed that there is a dissociation between these abilities and success in related manipulation tasks—in human infants, sensitivity to launching and object permanence emerges before the ability to use such information to manipulate and a similar dissociation appears to be present even in non-human adult primates. From an interventionist perspective, this raises the very interesting question of the relationship between these two sets of abilities—is it appropriate to think of the abilities associated with sensitivity to launching and to object recognition as manifesting causal understanding at all, if these do not transfer to capacities for manipulation and control? How good is the evidence for dissociation/non-transference? What are the processes by which, at least in older children, such transfer is achieved, and what implications does this have for how we should think about what it means to possess a concept of causation?

Finally, it is a striking fact that other primates, including chimps, are greatly inferior to humans, including small children, in causal understanding, particularly in connection with tool use and object manipulation. The source and character of these deficits is a matter of ongoing controversy but one possibility, suggested by the primatologists Call and Tomasello (1997), and by Woodward (2007*b*), is that non-human primates possess only an egocentric kind of causal (or cause-like) understanding—they readily learn about the instrumental consequences of their own actions but fail to appreciate that the very same causal relationships can be present both between their own actions and their effects, between the actions of conspecifics

and the outcomes of their actions and between events occurring in nature that do not involve the actions of other creatures at all. One indication of this limitation is the apparent difficulty that non-human primates have in transferring information across these different contexts: for example, they don't seem very good at learning what the consequences would be if they were to perform various actions by observing the consequences of the actions of others or by passive observation of causal relationships as they occur in nature. (This in turn is connected to the well-known limitations of non-human primates in tasks involving imitation.) By contrast, even very small human children are adept at such learning. These observations fit naturally with the picture of the relationship between causal understanding and the outcome of interventions suggested in sect. 4—the human notion of causation transcends the actor's own experience of agency and is rather the notion of a relationship that has to do with what would happen under an abstract notion of intervention that can be realized by other actors or by nature.

FURTHER READING

Menzies and Price (1993) is the most philosophically sophisticated recent defence of an agency theory. Hausman (1998) is a very detailed and systematic exploration of the interrelations between manipulation, agency, and causal asymmetries. Pearl (2000) is a lucid presentation of a broadly manipulationist approach to causation within a Bayes net framework, with emphasis on the formal characterization of various causal notions, and their representation in terms of directed graphs. Both Pearl and Spirtes, Glymour, and Scheines (1993) discuss and motivate the arrow-breaking conception of intervention and the prediction of the effects of interventions. The latter particularly focuses on issues of causal inference but both books are philosophically very rich, as well as important contributions to the allied literature in statistics and artificial intelligence. Finally, Woodward (2003) develops an interventionist approach to causation and explanation.

REFERENCES

CALL, J., and TOMASELLO, M. (1997). *Primate Cognition*. New York: Oxford University Press.

CAMPBELL, D. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

CARTWRIGHT, N. (1983). *How the Laws of Physics Lie*. Oxford: Clarendon.

COLLINGWOOD, R. (1940). *An Essay on Metaphysics*. Oxford: Clarendon.

COOK, T., and FRISCH, M. (2002). 'Non-Locality in Classical Electrodynamics', *British Journal for the Philosophy of Science* 53: 1–19.

GASKING, D. (1955). 'Causation and Recipes', *Mind* 64: 479–87.

GLYMOUR, C. (2004). 'Review of James Woodward, *Making Things Happen: A Theory of Causal Explanation*', *British Journal for Philosophy of Science* 55: 779–90.

HALL, NED (2000). 'Causation and the Price of Transitivity', *Journal of Philosophy* 97: 198–222.

HALPERN, J., and PEARL, J. (2001). 'Causes and Explanations: A Structural-model

Approach—Part I: Causes', *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann, 194–202.

HAUSMAN, D. (1986). 'Causation and Experimentation', *American Philosophical Quarterly* 23: 143–54.

—— (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.

HITCHCOCK, C. (2001a). 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98: 273–99.

—— (2001b). 'A Tale of Two Effects', *Philosophical Review* 110: 361–96.

—— (2007a). 'What Russell Got Right', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press.

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

HOLLAND, P. (1986). 'Statistics and Causal Inference', *Journal of the American Statistical Association* 81: 945–60.

LESLIE, A., and KEEBLE, S. (1987). 'Do Six-Month-Old Infants Perceive Causality?' *Cognition* 25: 265–88.

LEWIS, D. (1973). 'Causation', *Journal of Philosophy* 70: 556–67.

MEEK, C., and GLYMOUR, C. (1994). 'Conditioning and Intervening', *British Journal for the Philosophy of Science* 45: 1001–21.

MENZIES, P., and PRICE, H. (1993). 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science* 44: 187–203.

NORTON, J. (2007). 'Causation as Folk Science', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press.

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

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

PRINZ, J. (2002). *Furnishing the Mind: Concepts and their Perceptual Basis*. Cambridge, Mass.: MIT.

RUBIN, D. (1986). 'Comment: Which Ifs Have Causal Answers?' *Journal of the American Statistical Association* 81: 961–2.

SCHAFFER, J. (2000). 'Causation by Disconnection', *Philosophy of Science* 67: 285–300.

SPIRTES, P., GLYMOUR, C., and SCHEINES, R. (1993). *Causation, Prediction and Search*. New York: Springer. 2nd edn. 2000. Cambridge, Mass.: MIT.

VON WRIGHT, G. (1971). *Explanation and Understanding*. Ithaca, NY: Cornell University Press.

WOODWARD, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

—— (2007a). 'Causation with a Human Face', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press.

—— (2007b). 'Interventionist Theories of Causation in Psychological Perspective', in A. Gopnik and L. Schulz (eds.), *Causal Learning: Psychology, Philosophy and*

Computation. New York: Oxford University Press.

—— and HITCHCOCK, C. (2003). 'Explanatory Generalizations, [Part 1](): A Counterfactual Account', *Noûs* 37: 1–24.

PART III
ALTERNATIVE APPROACHES TO CAUSATION

CHAPTER 12
CAUSAL POWERS AND CAPACITIES

STEPHEN MUMFORD

1. OVERVIEW

A dispositional ontology, admitting a category of power or capacity, is thought by some to offer a vital insight into the nature of causation. Proponents believe that other ontologies lack the metaphysical resources to capture this insight. At its most ambitious, a causal powers ontology purports to offer a solution to, or dissolution of, the problem of causation. The argument is that the traditional problem of causation is generated by a faulty Humean ontology in which the world is described as a sum of 'loose and separate' distinct existences (Hume 1748: 74). Once the main Humean premiss is accepted, of there being no necessary connections between distinct existences, then the notion of causation becomes immediately problematic. Does it really exist? Is it just a constant conjunction? Can anything cause anything else? How must events stand to count as causally related? Can we have causal and inductive knowledge? This most common sense and widely accepted of notions, which is arguably a part of everyone's pre-critical understanding of the world, is then either rejected outright or its defenders are forced to fight a rearguard action, presenting ever more elaborate theories to underwrite something so simple and basic. Dispositionalists, on the other hand, think that while the strategies detailed in Part II of this volume might have varying degrees of success in describing what is true when there is causation, they are not about the causation itself. The regularity theory and the counterfactual theory are judged merely to describe phenomena that follow from real causal connections. They are attempts to avoid acknowledging the reality of causal necessity. Such theories face a Euthyphro question. For example, is there causation because there is constant conjunction or is there constant conjunction because there is causation? The dispositionalists think there will be constant conjunctions, counterfactual or probabilistic dependencies, and so on, because there are real causal connections in nature.

Where it is most radical, the powers ontology proposes a major reconceptualization of causation. Hume, as traditionally interpreted, understood the world to consist of distinct and discrete, unconnected existences. If this is accepted, then the best that can be made of causation is that it is a contingent and external relation between such existences. The powers ontology accepts necessary connections in nature, in which the causal interactions of a thing, in virtue of its properties, can be essential to it. Instead of contingently related cause and effect, we have power and its manifestation, which remain distinct existences but with a necessary connection between.

The powers ontology faces difficulties of its own, however. The programme is still in its relative infancy and it needs to be shown in detail how the theory accounts for causation. In most cases, it has been assumed that a theory of causation falls unproblematically from the theory of powers, but it is not immediately obvious how it does so (and just calling them *causal* powers is no substitute for such an account). In addition to this, opponents ask what reason we have to believe in such things as causal powers in the first place. Dispositionalists do not rise to the Humean sceptical challenge about necessary connections in nature. They tend either to ignore it or say that it is not worth answering. Hume had asked what reason there was to believe in necessary connections in nature and he professed he could see none (Hume 1739–40: 1. 3. 14; 1748: §7). There is, nevertheless, optimism among dispositionalists that both these challenges can be met. In the first place, an account of how to get causes from powers was begun by Molnar (2003: ch. 12) and this account may yet come to be the foundation of a robust dispositionalist theory of causation. In the second place, dispositionalists regard Hume's challenge as a form of scepticism that cannot be tackled on its own terms. The powers ontology rejects the whole apparatus that raises causation as a problem. The argument for accepting powers is their overall productivity compared to Humean alternatives. Like other theories in metaphysics, this one is to be accepted because, its supporters argue, it accounts for troublesome phenomena in a plausible and economical way. If one accepts powers as the basic ontological category, one can produce plausible theories of properties, laws of nature, modality, and causation. The success of producing a theory of causation is, therefore, one of the reasons for accepting the powers ontology in the first place.

2. HISTORY

The causal powers ontology can still be considered new. Although powers and dispositions were discussed by the likes of Boyle ([1666] 1979), Broad (1925), Carnap (1936), and Bergmann (1955), who saw how often they were invoked in science, it was not until the 1960s and 1970s that philosophers started to push powers as the central commitment of a new metaphysic. This powers metaphysic comes from two distinct traditions that developed in relative isolation.

One such tradition was based in Britain and came from the work of Rom Harré (1970; 2001, and with Madden 1973; 1975), which seems to have been an influence on Roy Bhaskar (1975) and Nancy Cartwright (1983; 1989; 1999). In Cartwright, the commitment is to capacities, which in her account differ from dispositions in that they 'are not restricted to any single kind of manifestation ... [but] can behave very differently in different circumstances' (1999: 59). Added to this mainly British tradition, though not obviously influenced by it, is the work of Hugh Mellor (1974), who followed a discussion of dispositions that appeared in Popper's *Logic of Scientific Discovery* (1959: app. 10). It is true of all these contributions that they came largely from considerations in the philosophy of science and a recognition of how important the dispositional vocabulary was in scientific theories. Physics in particular seems to invoke powers, forces, and propensities, such as the spin, charge, mass, and radioactive decay of subatomic particles. It occurred, particularly to Harré and Bhaskar, that empiricist philosophers may have developed an ontology that was completely inappropriate to serve

science and might be a fundamental misconception of the way the world works.

Meanwhile, a distinct tradition developed in Australia, where the main considerations were metaphysical rather than scientific. The originator of this tradition was C. B. Martin, who in the 1960s gave a seminar on causation at the University of Sydney with George Molnar (Martin claims that such ideas originally came to him as early as 1957—see his 1994: 8n.). In the seminar series, Martin and Molnar developed their anti-Humean alternative account and although many of their ideas remained unpublished they nevertheless became known through the word of their students and associates. One such early paper, which was published only recently, was Martin (1994) in which he argued that dispositions could not be reduced away in terms of conditional statements and had, therefore, to be accepted as real. This view is supported and developed in his other papers (Martin 1984; 1993*a*; 1993*b*; 1996). Molnar's own work on powers was also kept out of the public domain by a lengthy break from academia, but finally it has surfaced (1999; 2003). Also recently, Brian Ellis has produced his version of the anti-Humean metaphysic, which he calls dispositional essentialism (2001; 2002; also with Lierse 1994). This work contains a detailed analysis and critique of the Humean ontology of distinct existences.

There is now at last a cross-fertilization of the British and Australian traditions, shown in the work of Mumford (1998; 2004), Handfield (2001; 2005) and many others, which is both scientifically and metaphysically informed. There is also a late but growing interest in America, for instance by McKitrick (2003; 2005), Cross (2005), and long-time associate of C. B. Martin, John Heil (2005 and Martin and Heil 1998). The powers metaphysic is getting a more detailed examination than it has ever before.

3. WHAT IS A POWER?

Dispositions were, until recently, conceived of as a kind of property. Properties were thought to divide into two kinds: dispositional and categorical (as representative of this view, see Prior 1985, but also Armstrong 2005). This division is no longer so popular among those who support the anti-Humean ontology. For one thing, powers theorists are suspicious of the notion of the categorical. If a dispositional property is supposed to be a power then a categorical property, presumably, would be powerless. But should we allow that something could be a property if it were powerless? This would seem to mean that it can have no effects, or none essentially, neither on observers nor on other properties. If this is the right way to understand categorical properties—as intrinsically powerless—then they sound like mere epiphenomena that make no real difference to the world. Is it not a more appealing view that any property must make some difference to the behaviour, actual or at least possible, of a thing that instantiates it? In that case, there would be no categorical properties. Shoemaker develops this view ([1980] 2003*a*), arguing that all properties should be understood as powers, or bundles of powers. Shoemaker subsequently abandoned this view ([1998] 2003*b*) but it is not clear that he was right to do so (Mumford 2008).

Dispositions might not be best thought of as a kind of property, therefore. But how should they be understood? The leading powers theorists now typically present powers as a distinct and basic ontological category in their own right, irreducible to any other ontological

category. Few philosophers wish to multiply entities beyond necessity but the rationale for adding a category of power is that a potentially economical ontology is in prospect. Acceptance of powers could lead to accounts of properties, laws of nature, modality, causation, and perhaps more. Many other metaphysical categories might be explained and unified. If it can be demonstrated that a metaphysical commitment can indeed produce this, it is regarded as a good reason for accepting such a commitment. The problem for the powers metaphysic, however, is that such work is still programmatic with nothing that is widely regarded as demonstrated.

Is there some positive characterization of powers that shows their important features? Powers are taken to be real but clearly they are not substance-like existents. Although powers are not properties, they are nevertheless taken to be universals that have their instances in substances. Like a number of other metaphysical categories, the type–token distinction is applicable to them. Hence solubility might have a token—a particular instantiation—in a certain sugar cube and be tokened in other things at the same time. There is identity across these instantiations, however, in that they are all of the same type. The type need not be some extra, transcendental object but could exist in its instantiations, as Armstrong's (1978) immanent realist theory of universals suggests. When we speak of solubility in general, we are speaking of the type, but this exists only in so far as there are tokens of that type instantiated by particular objects.

There are two further significant features in the characterization of powers. These mark the powers ontology clearly as anti-Humean. First, each power is essentially, or necessarily, related to manifestations of a specific kind. Hence, solubility is essentially related to dissolving; fragility is essentially related to breaking. Second, however, the power and its manifestation are distinct existences. The power can exist without being manifested, if it is not stimulated for instance. When these two features are combined, they have the clear result that there can be necessary connections between distinct existences—in direct contradiction to Humeanism.

Latter-day Humeans are dismissive of such a necessary connection and will almost certainly want, like the logical positivists, to treat all such necessity as merely analytic. They will say that a power's connection to its manifestation is only a conceptual truth and any resulting necessity is merely *de dicto* rather than, as powers realists want, *de re*. Further, if the connection between the power and its manifestation is a solely analytic one then it might not even be correct to consider them distinct existences. If either the first or further objection is valid, the Humean thesis remains unscathed.

3.1 Are Powers and their Manifestations Distinct?

Let us first look at the evidence for the power and its manifestation being distinct existences. A power, it is claimed, can be held even if it is not manifested. Arguably, powers can be held even if they are never manifested for all the time they are held. Typically, their manifestations occur only in specific circumstances, where they are stimulated, enabled, or released. Those circumstances might never arise; nevertheless, the manifestations remain real possibilities while ever the power is possessed. If an object has a power that is never manifested, there may be some scepticism from Humeans that it is really there. The powers

theorist sees this as a mistake springing largely from empiricist assumptions. The empiricist will want to cash out a power ascription in terms of observable events and will therefrom consider test situations for the power. However, the power may never be appropriately tested or, even when it is tested, there may be some interfering factor that prevents manifestation. In theory, this could happen each time a test is performed (Wright 1991). As noted by Mellor (1974: 167), there could be cases where the very test for the power directly prevents the power manifesting. His example is of a nuclear reactor that has the disposition to explode. When it is about to do so, a safety mechanism cuts in and shuts the reactor down. Is an empiricist to claim that because the reactor never does explode then it doesn't have the disposition to explode in the first place? Such an answer is counterintuitive because if the reactor did not have the disposition to explode then there would be no point in the safety mechanisms. The case suggests, therefore, that the structure of the world's events, while it may be a result of the underlying powers of things, does not reveal all those underlying powers. Bhaskar (1975: 229) seems to be conveying such a thought when he distinguishes events from generative mechanisms, 'which are understood as tendencies and powers of enduring and transfactually acting things'. How something can act 'transfactually' seems a mystery but in the final section I will present some cases where sense can be made of this.

A power is plausibly a distinct existence from its manifestation, therefore, but to add to the case we can consider some further points. Manifestations must be distinct from the powers because, for one thing, they are of a distinct category. Following Ryle (1949) they are usually thought of as occurrences or events whereas powers are more like enduring states. The manifestations must further be distinct because it is not necessary that the power be still possessed once it is manifested. Though some powers can be maintained through their manifestation, others are not. If a soluble substance is dissolved it would not, in that state, continue to be soluble.

3.2 Are Powers and their Manifestations Necessarily Connected?

We need next to consider the connection between powers and their manifestations and whether it is a true case of *de re* necessity in things or just *de dicto* necessity in words. The powers theorist is greatly aided in this by Kripke (1980), who argued that there were metaphysical necessities, known a posteriori. If there are such metaphysical necessities then it raises the prospect that they may come in other kinds, such as the necessary connections claimed by essentialists between natural kinds and their essential properties and, of course, causal necessities.

This relation of power to manifestation is, however, of a peculiar kind. Given that a power may be held—perhaps is typically held—unmanifested, then it must be said to bear a relation, allegedly one of necessity, to something that does not exist. But a relation is real only when both its relata are real. Here one relatum is missing so how can there be any necessitation relation?

There are two main strategies for dealing with this problem. One is to understand dispositions in terms of counterfactual conditionals. If the power had been appropriately tested

then the manifestation would occur. This approach is not popular among powers theorists, however. The attempt to analyse dispositions into conditionals was primarily an empiricist strategy for converting claims about dispositions into claims about observable events, as evinced in Ryle (1949). The point of a powers ontology, however, is precisely that such a translation cannot be performed. Martin produced a purported proof of this (1994). A wire is live, meaning that it has a power to pass current to a conductor. But the wire is attached to an electro-fink, which is able instantaneously to render the live wire dead when it is touched by a conductor. Although live, it is false of the wire that 'if touched by a conductor, then current flows'. The fink can also work on a reverse cycle, rendering a dead wire live whenever it is touched by a conductor. Here the disposition is not possessed but the corresponding conditional is true. There has been continuing discussion of finkish and other problem cases (see Lewis [1997] 1999; Bird 1998; Choi 2005) and it is clear that Humeans have not yet given up on the conditional analysis of dispositions. Regardless of the finkish cases, however, Molnar (2003: 85–7) has further challenged the connection between dispositions and conditionals suggesting that there can be powers that are manifested spontaneously or continuously and for which rendition in conditionals seems entirely otiose. The point, however, is that for realists about powers the conditional analysis is largely rejected so they are not likely to invoke it to explain the relation between a power and its non-existent manifestation.

Another attempt to capture the connection is in terms of directedness. In the physical intentionality accounts of Molnar (2003: ch. 3) and Place (1996*a*; 1996*b*) dispositions are directed towards their manifestations just as mental states can be directed. The manifestation is the intentional object of the power. In the case of an unmanifested power, there is directedness towards an intentionally inexistent object. Intentionality, in this account, is said to be the mark of the dispositional rather than the mark of the mental, though mental states still qualify as intentional where they are also dispositional. This theory is a relatively recent one, which has yet to attract many followers. The chief reservation seems to concern whether a purely physical power can accurately be described as *directed* towards its manifestation (see Mumford 1999 and Place's reply 1999) or whether such language is nothing more than a metaphorical way of gesturing towards something else.

Perhaps that something else is causal necessity. If it is essential to a power that it is a *cause* of some type of manifestation, when the conditions are appropriate, then this might fit the bill. Such a relation is almost certain to be one that holds between the universals—the types—and instantiated in its tokens. It is being dissolved, qua universal, to which solubility is essentially related. It does not matter, in that case, that its particular instantiation is not yet existent. Its instantiation will have many peculiar features, being at some time and place, but it is not the peculiar instantiation of solubility to which solubility is related, it is dissolved-ness in general. What we would have, therefore, would be a necessary connection between universals. This looks more like a causal relation where, for example, it is plausible that if event *a* causes event *b* it does so in virtue of the properties of *a* and *b*; that is, causal efficacy is in virtue of properties. The chief causal relations are thus at the level of universals with individual causal processes occurring in virtue of the universals instantiated in the particular cause and effect.

4. How to Get Causes from Powers

We have seen what is meant by a power and some of the difficulties that surround the notion. A model of the relation between powers and their manifestations has been proposed in which it is primarily a causal relation, with powers waiting to be released or stimulated into action: 'ready to go' as Martin (1993: 180) has said. But what is the exact relation of powers to causes? How do we get causes from powers? Does *a* cause *b* simply when *a* has a power to cause *b*?

The simple answer is that each event that occurs can be regarded as an effect of a power manifesting itself in a causal process. In the appropriate conditions, a power will be a cause of its manifestation and the manifestation is the effect of the power. Ellis (2002: 48) presents such a view when he says a causal power is 'a disposition to engage in a certain kind of process: a causal process. A causal process is one of a kind that relates two events'.

At this point, a further move is possible. This is the additional claim that causation itself is nothing more nor less than the manifestation of powers. One distinct advantage of such a view is that it respects the singularist intuition that whether one token event *a* causes another token event *b* depends on nothing more than *a* and *b* and any relation between them. Singular causes can, in this ontology, come first with any general causal claims supervenient on them. Cartwright (1989: 141) at one point endorses something like this view when she says 'the most general causal claims—like "aspirins relieve headaches" or "electromagnetic forces cause motions perpendicular to the line of action"—are best rendered as ascriptions of capacity'. Molnar also follows Tooley's (1987) emphasis on the primacy of singular causation, though accepting that this view is logically independent of a powers ontology. Although Cartwright (2007: 11–23) places singular causal claims first, and sees them as dependent on capacities, she does not, however, offer this as a reductive analysis of causation. Rather, she believes that causation is a heterogeneous concept. A more orthodox view is that causation can be understood as itself a universal, existing immanently in its instances, which means that it is exactly the same thing in each case (Molnar 2003: 187–8). It is instantiated whenever a power is a cause of a manifestation, which requires that there is a *de re* necessity between the power and manifestation-type.

However, it is unlikely that the relation of powers to causes is quite so simple as this because it is implausible that every event is the manifestation of a power. Different powers work with each other or sometimes against each other to produce the history of events. This is frequently acknowledged by powers theorists. Cartwright (1983: 11–18), for instance, makes use of Mill's (1843: Bk. 3 ch. 6) notion of the composition of causes. Various powers would then each add their effect to produce a larger combined effect. Each spectator at a sporting event, for example, adds their small voice to the whole to produce a roar that none of them could have produced alone. Similarly, members of a choir each produce a single note with the combined effect of the whole producing a chord. Ellis (2002: 50) endorses this view when he says 'processes have a way of occurring one on top of another in a kind of avalanche, obscuring each other's effects'.

Molnar (2003: 186–99) has produced a detailed account of composition of causes in relation to the manifestation of powers. Events are typically polygenic: they are produced by many powers with small additive effects combining together. His illustration is a case of two horses pulling a barge from opposite banks of the canal. One horse exerts a force south-west. The

other horse exerts an equal force northwest. The barge moves in neither of these directions but instead follows a straight course west, down the centre of the waterway. Powers, in Molnar's classification, are pleiotropic: they make a contribution to many different effects. How, then, can a power be essentially related to its effect? Molnar asserts, in contrast to Cartwright's capacities, that the same power always makes exactly the same contribution to an effect, even though some other powers may prevent that contribution receiving an unimpeded manifestation. This suggests that a further division is being made. A power's manifestation is now characterized as its *contribution* to an effect event and not usually as an event itself. He sums his position thus: 'Manifestations are isomorphic with powers because each power gets its identity from its manifestation. ... A manifestation is typically a *contribution* to an effect, an effect is typically a *combination* of contributory manifestations. In other words, events are usually related as effects to a collection of interacting powers' (ibid. 195).

This makes the account of causation more complicated but arguably more plausible. It suggests, among other things, that some of the standard examples of powers, used to introduce the concept, are oversimplifications. Rarely will we see an event that is a manifestation of a single power. Instead we are more likely to see effects that are the outcome of composed manifestations. As Cartwright (1999: 46–7) notes, to isolate the manifestation of a single power, we may have to resort to the artificial conditions of the laboratory, where all other combined and interfering powers are screened off.

It is claimed by Molnar (2003: 188–90) that such an account has the advantage over other popular theories of causation. It respects the appeal of singularism, which the regularity theory does not. It is objectivist, which manipulability accounts of causation are not. It treats causation as a metaphysically real type so it does not face the sort of Euthyphro questions that can be asked of the counterfactual theory and probability-raiser theory. Lewis ([1973] 1986) may be right that there are relations of counterfactual dependence involved in causation. For Lewis, the Humean, such relations constitute causation. The realist about powers need not deny that such counterfactual dependencies hold but they are likely to view them as symptomatic of causation, not constitutive of causation. Similarly for the probability-raiser theory. That one event raises the chance of another is a symptom of a causal relation holding between the two. It does not constitute exhaustively the causal relation. The theory of causal powers is an attempt to say what causality actually is. It is not, as we might think of some other theories, a mere description of phenomena that accompany causality.

5. RECONCEIVING CAUSES

I said at the beginning that, at its most ambitious, a powers ontology might offer a dissolution of the problem of causation. I had in mind the idea that the problem of accounting for causation is one generated by Hume's emaciated metaphysic and that if one replaces it with a metaphysic of powers, the problem of causation simply goes away.

Such an idea is suggested by Ellis (2001: ch. 7) though acknowledged by many anti-Humeans. The argument is that the traditional way of conceiving causation springs largely from the underlying Humean ontology of unconnected events. This has handicapped subsequent discussion, as even many of those who would like to be causal realists nevertheless

retain an underlying Humean ontology. Some (Ellis has in mind Armstrong 1983) preserve the old Humean framework and attempt to impose necessity on top. But this is still a contingent and external relation (*contingent necessitation,* in Armstrong) between distinct existences. It is still, as Harré and Madden (1975: 89) say, 'a picture of nature as a crowd of passive sufferers of external and imposed causality'.

Humeanism has a metaphysic of discrete, distinct existences, usually understood to be events, which are self-contained. As Hume says, all is 'loose and separate'. This is the metaphysics of *discreta* (Mumford 2004: 182–4). If this picture of reality is accepted, causation can then only be understood as an asymmetric, dyadic, and external relation between such distinct existences. As Hume would say, there is nothing in *a* or *b* that makes them bear the causal relation so, if the relation exists at all, it must be something additional to them (this is what is meant by an external relation). Furthermore, such a relation has to be contingent because, if it is an external relation between discrete existences, anything can in principle be causally related to anything else. Humeanism then leads us towards causal scepticism. With the empiricist strictures on admissible evidence, we are said to experience only the discreta and cannot experience any necessary connection between them. This is why Hume says that the only legitimate idea of causation is constant conjunction (plus contiguity and temporal priority). But constant conjunction allows no sense of causation in the single case. This is only the traditional account of Hume, however. Strawson (1989) would have it that Hume was really a causal realist, believing in necessary but secret connections. Strawson makes a good case for a metaphysics of causal powers though there remains some doubt that he has pinned the position on Hume (see Mumford 2004: 57–61).

Standard examples are used that specially fit the discrete model: billiard balls crashing into each other, for instance, in clearly distinguished, self-contained events (Hume 1739–40: 164). But does the world neatly divide into causes and effects with an asymmetric causal relation between them? Different examples would not fit the Humean paradigm so easily: two books leaning against each other, each keeping the other from falling; planets orbiting their stars in equilibrium, maintaining a stable distance apart; a fridge-magnet sitting motionless in place. Forces are 'acting' in all these cases though, in a sense, nothing is happening. These forces 'act' transfactually, as Bhaskar would say, but they do so to prevent paradigmatic Humean events from occurring, rather than bringing them about. If in doubt that such forces are at work, take away one book and watch the other fall, take away the sun and watch the planet fly away into the ether, introduce a distance between the magnet and the fridge door and feel the force of the attraction fade. Until that is done, however, nothing much is happening that would neatly qualify as a Humean event. We have non-events because of the composition of powers to create a static effect.

If these are acceptable cases of causation, then there is significance in them. As well as an absence of Humean events, there is no asymmetric causation, no temporal priority, no clear distinction between cause and effect. Instead we have simultaneous forces acting to preserve an equilibrium state. Martin (1993: 516*a*) speaks of 'reciprocal mutual dependence' in 'the production, prevention or the *continuance* and *sustaining* of various properties of an entity', which fits better the new examples of causation. Reciprocity of cause and effect is a principle of causation often accepted in science and metaphysics, though the philosophical implications remain controversial (see Le Poidevin 1988).

We might now have arrived at the proposed reconceptualization of causation. It need no longer be an asymmetric, external relation between discreta. Instead there is reciprocity and causation between distinct existences. In such a case, the causal relation would not even be an external one. The existence of the causal relata, the power and its manifestation, is enough alone to ensure that the causal relation exists. Such a relation is called an internal relation. The essence of a power is that it be connected to such manifestations. Hence nothing could have mass without attracting other things with mass in the way described by the gravitation law. A thing could not have mass and be loose and separate from the rest of the world. Finally, it is worth mentioning that the relata of causal relations will no longer primarily be events at all. They will primarily be the powers and properties of things, the universals that particular objects instantiate. If all these features could be developed in a new theory of causation, it would amount to a radical rethinking of the issue from the way Hume originally introduced it.

It is worth repeating that many of these claims are yet to be fully developed and that there is ample room for further investigation within this framework. A number of metaphysicians are sufficiently encouraged by the progress to date to consider further work worthwhile.

FURTHER READING

The classic account of causal powers within the philosophy of science tradition is Harré and Madden (1975). For an attack on the alternative, empiricist view of science see Bhaskar (1975), and, for something more technical, Cartwright (1999). In metaphysics, the powers ontology is developed by Molnar (2003). Ellis (2001) presents powers with essentialism as a radical alternative to Humean metaphysics, and Mumford (2004) takes this line further, arguing that the old framework of properties and laws is inadequate.

REFERENCES

ARMSTRONG, D. M. (1978). *A Theory of Universals*. Cambridge: Cambridge University Press.
—— (1983). *What is a Law of Nature?* Cambridge: Cambridge University Press.
—— (2005). 'Four Disputes About Properties', *Synthese* 144: 309–20.
BERGMANN, G. (1955). 'Dispositional Properties and Dispositions', *Philosophical Studies* 6: 77–80.
BHASKAR, R. (1975). *A Realist Theory of Science*. Leeds: Leeds Books.
BIRD, A. (1998). 'Dispositions and Antidotes', *Philosophical Quarterly* 48: 227–34.
BOYLE, R. ([1666] 1979). *The Origin and Forms and Qualities*, in M. A. Stewart (ed.), *Selected Philosophical Papers of Robert Boyle*. Manchester: Manchester University Press, 1–96.
BROAD, C. D. (1925). *The Mind and its Place in Nature*. London: Harcourt Brace.
CARNAP, R. (1936). 'Testability and Meaning', *Philosophy of Science* 3: 420–71.
CARTWRIGHT, N. (1983). *How the Laws of Physics Lie*. Oxford: Clarendon.
—— (1989). *Nature's Capacities and Their Measurement*. Oxford: Clarendon.
—— (1999). *The Dappled World*. Cambridge: Cambridge University Press.
—— (2007). *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

CHOI, S. (2005). 'Do Categorical Ascriptions Entail Counterfactual Conditionals?', *The Philosophical Quarterly* 55: 495–503.

CROSS, T. (2005). 'What is a Disposition?', *Synthese* 144: 321–41.

ELLIS, B. (2001). *Scientific Essentialism.* Cambridge: Cambridge University Press.

—— (2002) *The Philosophy of Nature: A Guide to the New Essentialism*. Chesham: Acumen.

—— and Lierse, C. (1994). 'Dispositional Essentialism', *Australasian Journal of Philosophy* 72: 27–45.

HANDFIELD, T. (2001). 'Dispositional Essentialism and the Possibility of a Law-abiding Miracle', *The Philosophical Quarterly* 51: 484–94.

—— (2005). 'Armstrong and the Modal Inversion of Dispositions', *The Philosophical Quarterly* 55: 452–61.

HARRÉ, R. (1970). 'Powers', *British Journal for the Philosophy of Science* 21: 81–101.

—— (2001). 'Active Powers and Powerful Actors', *Philosophy* 76 suppl.: 91–109.

HARRÉ, R., and MADDEN, E. H. (1973). 'Natural Powers and Powerful Natures', *Philosophy* 48: 209–30.

—— (1975). *Causal Powers*. Oxford: Oxford University Press.

HEIL, J. (2005). 'Disposition', *Synthese* 144: 343–56.

HUME, D. (1739-40). *A Treatise of Human Nature*, ed. L. A. Selby-Bigge. Oxford: Clarendon, 1888.

—— (1748). *An Enquiry Concerning Human Understanding*, in *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L. A. Selby-Bigge. 3rd edn. rev. P. H. Nidditch. Oxford: Clarendon, 1975.

KRIPKE, S. A. (1980). *Naming and Necessity*. Oxford: Blackwell.

LE POIDEVIN, R. (1988). 'The Principle of Reciprocity and a Proof of the Non-simultaneity of Cause and Effect', *Ratio* NS 1: 152–62.

LEWIS, D. ([1973] 1986). 'Causation', *Journal of Philosophy* 70: 556–67; also in *Philosophical Papers*. Oxford: Oxford University Press, ii. 159–72.

—— ([1997] 1999). 'Finkish Dispositions', *Philosophical Quarterly* 47: 143–58; also in *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press, 133–51.

MCKITRICK, J. (2003). 'The Bare Metaphysical Possibility of Bare Dispositions', *Philosophy and Phenomenological Research* 66: 349–69.

—— (2005). 'Are Dispositions Causally Relevant?', *Synthese* 144: 357–71.

MARTIN, C. B. (1984). 'Anti-Realism and the World's Undoing', *Pacific Philosophical Quarterly* 65: 3–20.

—— (1993*a*). 'The Need for Ontology: Some Choices', *Philosophy* 68: 505–22.

—— (1993*b*). 'Power for Realists', in J. Heil (ed.), *Ontology, Causality and Mind*. Dordrecht: Kluwer, 175–86.

—— (1994). 'Dispositions and Conditionals', *Philosophical Quarterly* 44: 1–8.

—— (1996). 'Properties and Dispositions', in D. M. Armstrong, C. B. Martin, and U. T. Place, *Dispositions: A Debate*. London: Routledge, 71–87.

—— and HEIL, J. (1998). 'Rules and Powers', *Philosophical Perspectives* 12: 283–312.

MELLOR, D. H. (1974). 'In Defense of Dispositions', *Philosophical Review* 83: 157–81.

MILL, J. S. (1843). *A System of Logic*. London: Parker.

MOLNAR, G. (1999). 'Are Dispositions Reducible?', *Philosophical Quarterly* 49: 1–17.

—— (2003). *Powers: A Study in Metaphysics,* ed. S. Mumford. Oxford: Oxford University Press.

MUMFORD, S. (1998). *Dispositions*. Oxford: Oxford University Press.

—— (1999). 'Intentionality and The Physical', *Philosophical Quarterly* 49: 215–25.

—— (2004). *Laws in Nature*. Oxford: Routledge.

—— (2008). 'Powers, Dispositions, Properties', in R. Groff (ed.), *Revitalizing Causality: Realism about Causality in Philosophy and Social Science*. Abingdon: Routledge.

PLACE, U. T. (1996*a*). 'Intentionality as the Mark of the Dispositional', *Dialectica* 50: 91–120.

—— (1996*b*). 'Dispositions as Intentional States', in D. M. Armstrong, C. B. Martin, and U. T. Place, *Dispositions: A Debate*. London: Routledge, 19–32.

—— (1999). 'Intentionality and The Physical: a Reply to Mumford', *Philosophical Quarterly* 49: 225–31.

POPPER, K. R. (1959). *The Logic of Scientific Discovery*. Rev. impression 1980. London, Hutchinson.

PRIOR, E. W. (1985). *Dispositions*. Aberdeen: Aberdeen University Press.

RYLE, G. (1949). *The Concept of Mind*. London: Hutchinson.

SHOEMAKER, S. ([1980] 2003*a*). 'Causality and Properties', in P. van Inwagen (ed.), *Time and Cause*. Dordrecht, Reidel, 109–35; also in *Identity, Cause, and Mind*. Expanded edn. Oxford: Clarendon, 206–33.

—— ([1998] 2003*b*). 'Causal and Metaphysical Necessity', *Pacific Philosophical Quarterly* 79: 59–77; also in *Identity, Cause and Mind*. Expanded edn. Oxford: Oxford University Press, 407–26.

STRAWSON, G. (1989). *The Secret Connexion: Causation, Realism, and David Hume*. Oxford: Clarendon.

TOOLEY, M. (1987). *Causation: A Realist Approach*. Oxford: Oxford University Press.

WRIGHT, A. (1991). 'Dispositions, Anti-Realism and Empiricism', *Proceedings of the Aristotelian Society* 91: 39–59.

CHAPTER 13
ANTI-REDUCTIONISM

JOHN W. CARROLL

1. WHAT IS ANTI-REDUCTIONISM?

Philosophers routinely seek a certain sort of analysis of causation. They have sought a completion of

(S) *c* caused *e* if and only if ...

showing what makes causal facts both true and accessible enough for us to have the knowledge of them that we ordinarily take ourselves to have.

Some current approaches to analysing causation were once resisted. First, analyses that use the counterfactual conditional were viewed with suspicion because philosophers also sought (and still do seek) similar understanding of counterfactual facts. Since the same can be said for the other *nomic* concepts—causation, lawhood, explanation, chance, dispositions, and their conceptual kin—philosophy demonstrated a preference for non-nomic definitions of causation, analytic completions of (S) with no nomic terms in the analysans. Recently, however, philosophers have been less demanding regarding what terms may be used. Attention has been given to analysing causation in terms of chance, the counterfactual conditional, and lawhood. If we reserve the term 'causal' for the terms and concepts that have extremely obvious connections with causation (that is, causation itself and its close nomic cousins, for example, production, bringing about, and explanation), we can say that, of late, philosophers have only demanded that (S) be completed using solely *non-causal* terms. Second, philosophers once insisted that the completion of (S) be analytic, that it be a definition of the verb 'to cause'. Recently, however, they have only demanded that the analysis be a necessary truth. Some even hold that, so long as it can be maintained that the causal supervenes on the non-causal, concerns about the truth-makers for and our knowledge of causal facts would have been addressed.

Anti-reductionism is the view that causation cannot be analysed non-nomically and, further,

that causation still resists analysis even when the non-causal, nomic concepts are made available. In other words, the anti-reductionist maintains that there can be no non-causal analysis of causation. Indeed, some anti-reductionists hold that causation does not supervene on the non-causal facts. This chapter is an overview and defence of anti-reductionism. Section 1 is nearly complete. Section 2 locates anti-reductionism relative to some possible companion doctrines. Section 3 recounts the development of anti-reductionism. Arguments in favour of anti-reductionism are advanced in sect. 4. Anti-reductionism and its supporting arguments are defended against objections in sect. 5, before the chapter concludes in sect. 6 with some ruminations about current and future work on causation.

2. Related Doctrines

2.1 Reducibility of Higher-Level Causation

The antithesis of anti-reductionism is reductionism, the view that there could be a necessarily true completion of (S) using only non-causal terms in the analysans. Reductionism, however, needs to be distinguished from a view that goes by that same name or its minor variant *causal reductionism* (see Ch. 30). This other view might more revealingly be called *the reducibility of higher-level causation to lower-level causation*. It is the view that the causal claims of all higher-level sciences reduce to the causal claims of fundamental physics.

2.2 Singularism

It is standard to contrast single-case causal sentences ('the eruption of Mt. Vesuvius caused Pompeii's destruction') with what are sometimes called *general-case* or *property-level* causal sentences ('smoking causes cancer'). This distinction is important because frequently there is talk of *singularism* accompanying anti-reductionism even though several different doctrines go by that name. For example, it is natural to take singularism to maintain that general-case causal claims are not conceptually prior to single-case causal claims (cf. Cartwright 1989: 91–140; Carroll 1991; Lewis [1973] 1986: 161–2). Some take singularism to hold that there could be causation in a world devoid of laws of nature and other uniformities (cf. Ducasse [1926] 1993: 129). Menzies (1999: 315) and Armstrong ([2001]2004: 452–3) take singularism to include the view that the causal relation is an intrinsic relation. None of these forms of singularism entail or are entailed by anti-reductionism.

2.3 Primitivism

Sometimes anti-reductionism is labelled *primitivism*. That label, however, is better used for the view that causation is primitive. Such a view, unlike anti-reductionism, denies that there are *any* concepts more basic than causation. So, primitivism denies that the nature of causation can be revealed by an analysis of causation in terms of, say, explanation (e.g. Scriven 1975) or the bringing about relation (Gotshalk 1931). Furthermore, unlike anti-

reductionism, primitivism is bound to have some implications about our acquisition of the concept of causation (maybe that it is innate or arrives perceptually as a Humean impression) or about the epistemology of that concept (maybe that causation is knowable a priori or is directly experienced). Primitivism entails, but is not an entailment of, anti-reductionism.

2.4 Experience of Causation

Though it need not be, anti-reductionism is often associated with the view that we have direct perceptual or introspective access to causation. A weak version of this doctrine holds that some causal truths are trivially inferred from observable facts (Anscombe ([1971] 1981: 137). For example, someone might see Marvin hit Tommy and thereby know that Marvin hit Tommy. From such knowledge, it could then be inferred that Marvin caused a change to Tommy. But the thesis that we experience causation takes other stronger forms. Some want to hold that there is something like an impression of causation, a causation sensation (Armstrong 1993). Usually the pertinent kind of experience is claimed either to be an inward sensation associated with the initiation of action or else the tactile sensation of pressure. (For further discussion, see Ch. 22.)

2.5 Pluralism, Scientific Essentialism, and Anti-Realism

This chapter is built on certain (plausible) assumptions. For example, it is assumed that there is a primary sense of 'to cause'. *Pluralists* (see Ch. 16) argue that there are multiple causal relations, sometimes claiming that none is more central or primary or philosophically important than another. For a second example, it is assumed that causation is a contingent relation—a cause could exist even if none of its effects existed. *Scientific essentialists* (for example Ellis 2002; also see Ch. 12) hold (1) that causation is the manifestation of the powers of objects, and (2) that, if a cause occurs, its effects occur as a matter of necessity. For a third example, it is assumed in this chapter that *realism* is true about causation, that there are some causal sentences that purport to, and succeed in, describing reality. Some anti-realists, *the eliminativists* (Russell 1912–13), hold that these sentences don't succeed in describing the world, and so also hold that, strictly speaking, nothing causes anything else. Others, *the projectivists* (Blackburn 1990; Price 2007) will utter such sentences as 'the spark caused the fire', but are anti-realists in virtue of thinking that such utterances will project something about us rather than convey information about the way the world is independent of us.

2.6 Ostension, Theoretical Specification, and Theoretical Analysis

It is important to contrast three different ways one might complete (S) in a non-causal fashion, and still not obviously be in conflict with anti-reductionism.

(a) *Ostension*. Suppose that we sometimes have experiences of causation. Arguably, this would permit an ostensive specification of the causal relation. We could feel pressure from a

strong wind, and specify that what we felt was causation. Once the relation was specified, it could be referred to in an analysis without using 'to cause' or any other causal terms. Think of the ostension as introducing a proper name for the causal relation, say 'C'. Then one could give the following analysis: *c* caused *e* if and only if *c* and *e* (in that order) instantiated C. 'C' is not a causal term. If the ostension was successful, if there really was a relation assigned to 'C' and it really was the causal relation, this completion of (S) would be true. Menzies and Price (1993: 194–5) adopt a version of this approach.

(b) *Theoretical specification*. The specification need not be ostensive. Rather than pointing to the causal relation, the method of theoretical specification tries to pick it out theoretically. Suppose one had put together a theory consisting of some simple plausible claims about causation, claims such as that causation is an intrinsic and transitive relation, that often causes stand to effects as means stand to ends, and so on. One could specify that causation is the relation that makes the truisms true and then use a name to refer to that relation in the analysis (see Menzies 1996: 98–101; Armstrong [2001] 2004: 453–5).

(c) *Theoretical Analysis*. A third option is to incorporate the theoretical specification right into the completion of (S). One might maintain that *c* caused *e* if and only if *c* and *e* (in that order) instantiate the relation that satisfies the truisms (see Tooley 1984; 1987; 1990*a*; 1990*b*; 2003). This approach could turn out to be only a notational variant of the theoretical-specification approach. Whether it does so depends on how the relation that the specification *actually* picks out compares to the relation that would be picked out *in other possible worlds*. (Remember that analyses are expected to at least be necessary truths.) If the specification picks out the same relation in all possible worlds, then there is no difference of consequence between the two approaches. A rigid designator like the name 'C' can't help but pick out the same relation in all worlds.

Indeed, these three reductive approaches might be essentially equivalent to each other and not interestingly different from anti-reductionism. There would be significant agreement between the three manners of reducing causation if the relation picked out by the ostensive specification is the same relation as the one picked out by the theoretical specification, and the theoretical specification picks out the same relation in all possible worlds. There would be significant agreement with anti-reductionism if, in addition, the relation picked out by the ostension and the theoretical specification was *the very same relation that the anti-reductionist takes to be irreducible*. This potential agreement may explain why Armstrong, Menzies, Tooley, and the anti-reductionists sometimes employ similar arguments. Keep in mind, however, that an actual accord with anti-reductionism depends on the success of the ostensive or theoretical specification. Neither the ostension nor the decision about which of the so-called truisms to include in a theoretical specification is a trivial matter. If it turned out that the preferred attempt at specification picked out *nothing* or picked out the *wrong* relation —say, the is-adjacent-to relation—then all three of these reductive approaches would lose their appeal though anti-reductionism would be unblemished.

3. The Development of Anti-Reductionism

In the late nineteenth and early twentieth enturies, there are statements of anti-reductionism recognizable as the precursors to contemporary anti-reductionisms. So, for

example, Peterson (1898: 61) concludes:

If now the reader asks me what causation is, I reply that I think it is a simple, unanalyzable relation, not derived from anything nor resolvable into anything else. ... Of course, any man is free to analyze the relation if he can, but it is not likely that any one hereafter will succeed where thinkers so able as Hume and Mill conspicuously failed.

For a second example, there is Lamprecht. After providing numerous examples supporting 'the empirical status of necessity or compulsion in events' (1929: 193), he goes on to say:

We can state the genus of causality: it is a relation. But we cannot give the essential difference of the causal relation except in some question-begging synonym. We can say that causality is a *necessary* relation between cause and effect, or that it is the character of the process in which one thing *produces* another, or that it is *efficacious* control of one thing over another. These assertions are true; but they are not adequate as formal definitions. They do not advance the discussion one whit; they would not explain causality to any one who did not already know what we were talking about. (ibid.; also see Lamprecht 1930)

Embracing the non-reductive element of his own analysis, Gotshalk (1931: 475) points out that 'every formal definition of a term is in the end nothing but this term joined to a set of terms which have a meaning more or less equivalent to the essential meaning of the original' (also see Gotshalk 1930: 241). Broad (1925: 453–6) also shows sympathy to anti-reductionism.

Perhaps due to the influence of positivism, clear and forceful statements of anti-reductionism are hard to find from the mid-1930s until the mid-1960s. But, from the mid-1960s through the mid-1970s, we find the views of Scriven ('We can explain the relation between causal and non-causal language, but not by showing that one is built out of the other': 1966: 241, also see 239; and 1971: 51), Taylor ('To say of anything, then, that it was the cause of something else, means simply and solely that it was the cause of the thing in question, and there is absolutely no other conceptually clearer way of putting the matter except by the introduction of near synonyms for causation': 1966: 40), and Brand ('I shall consider the case of causation and show how in a programmatic way, a non-reductive analysis of causation is required': 1975: 151; also see Brand 1976). Anscombe ([1971] 1981: 137) endorses anti-reductionism in a paragraph where she defends the observability of causation: 'if we care to imagine languages in which no special causal concepts are represented, then no description of the use of a word in such languages will be able to present it as meaning *cause*'. During this period, the arguments for irreducibility began to be more sophisticated. Indeed, all the primary contemporary arguments for anti-reductionism are anticipated during these years (and will be cited when they are presented in sect. 4). Nevertheless, there was little development of the views beyond the rejection of extant analyses. There was also little reaction from the

reductionist community.

So, who are the contemporary anti-reductionists? Woodward (1990*a*; 1990*b*; 1994; 2003) and Carroll (1994; 2008) are examples. Their positions are in the spirit of Scriven, Taylor, Anscombe, and Brand in that there is not the appearance of providing a reduction. Like these earlier advocates, Woodward and Carroll present counterexamples to extant Humean/regularity analyses. But, more in line with reductionists such as Tooley and Armstrong, they explicitly and emphatically challenge supervenience. Woodward (1990*a*: 554, 557–9) and Carroll (1994: 161–81; 2008) distance themselves from Tooley and Armstrong by making clear their disagreement regarding the need for a reductive analysis of nomic facts. Neither Woodward nor Carroll is out to reveal the reductive truth-makers for causal facts.

4. ARGUMENTS FOR ANTI-REDUCTIONISM

4.1 Reductive Failures

A primary motivation for anti-reductionism is the repeated failures of reductive analyses, and rightfully so. That no successful non-causal analysis of causation exists ought to lead us to consider the possibility that there cannot be one. Many chapters of this handbook chronicle the attempted analyses and why they fail.

4.2 The Sparse Base

Consider perception. Though it is not counted as a causal concept, it has something to do with causation. The same goes for notions such as action, reference, and persistence. We might say that they are non-causal concepts that nevertheless do have *causal commitments*. The range of concepts with causal commitments is impressive: Nothing is a table unless it has a disposition to cause objects not to fall from its surface. Some have thought that colours are some sort of disposition to produce certain visual appearances. We have reasoned only if a judgement is caused by other of our mental states. For something to be material it must be impenetrable, it must be disposed to cause a sufficiently wide range of objects that may collide with it to be stopped without penetrating it.

From a reductionist perspective, because all these concepts do have causal commitments, they are ill-suited to provide a legitimizing analysis of causation. So, from this perspective, all the concepts with a causal commitment should be off limits. But are there any concepts free of causal commitment? Arguably, there are. The truth-functional concepts, standard mathematical concepts (for example, being prime), and logical necessity are examples. It may be that spatial and temporal relations lack causal commitment (see Ch. 20). This is more controversial because many are tempted to analyse temporal relations in causal terms. In any case, what is important to notice is that, even if we are generous by placing spatio-temporal relations in the class of concepts lacking causal commitment, this class still is quite barren. This is a serious problem for reductionism (cf., Scriven 1966: 240–1; Carroll 1994: 3–13). On the one hand, it is the distinction between the concepts with causal commitments and those

without, not the causal/non-causal or the nomic/nonnomic distinction, that is metaphysically significant. By the lights of the reductionist, only an analysis of causation that uses solely terms free of causal commitment should be acceptable. On the other hand, restrictions on the available vocabulary decrease the likelihood of success. Here, where we should be forced to restrict the vocabulary to terms free of causal commitment, it looks as if the likelihood of success is minuscule. The class of concepts that is truly autonomous, even if it is non-empty, is not rich enough to permit the desired analysis.

4.3 Directionality of Causation

Since Russell (1912–13: 13–14), philosophers have wondered what supplies the directionality that is evidently crucial to causation. Except in some special cases, causes precede their effects in time. For this same range of cases, causation is also an asymmetric relation. Russell believed that our world was deterministic, that the complete state of our world at any one time together with the laws of nature determined its state at all earlier times and at all later times. So, though he allowed that the laws of nature might account for a connection between two events, Russell thought there was nothing that could determine which of the two was the cause and which was the effect. Advances in physics question whether our universe is deterministic. But, as far as the question of reducibility is concerned, that does not matter. In order to succeed, an analysis needs to be necessarily true. So, the mere possibility of a deterministic world raises just as serious issues about the viability of a non-causal analysis.

It is tempting for the reductionist to put forward that time provides the requisite directionality by building into an analysis that c caused e only if the time of c is earlier than the time of e. Tooley (2003: 398) and others reject this move (and other temporal restrictions on causation) because it would compel one to be an antireductionist about time; with causation analysed in terms of time, there would then be no non-circular way of reducing directionality of time using only non-temporal terms. Another serious problem with this move is that it would rule out plausible cases of simultaneous causation. Suppose there is a perfectly rigid seesaw—when one end of the bar moves up or down, the other end moves in the opposite direction. You push down on one side. Then, it seems that your side simultaneously caused the other side to go up. The reductionist does no better by instead building into the analysis only that the time of c is not after the time of e. This weaker restriction does correctly say that the side you pushed down caused the other side to rise, but that is not enough. It incorrectly allows that the other side's going up caused your side to go down (see Taylor 1966: 35–40; von Wright 1971: 74–5; [1973] 1993: 118; 1974: 63–8; Carroll 1994: 141–7).

It is no objection to the seesaw example to point out that a perfectly rigid seesaw is physically impossible. It surely is: that there is a perfectly rigid seesaw contradicts the law that no signals travel faster than light. But all that really matters is that the seesaw case be possible. Also keep in mind that physicists take seriously the possibility of backwards-directed time-travel and the accompanying backwards-directed causation. It is taken as established that there are solutions to the equations of general relativity that include this sort of time-travel (see Gott 2001: 76–130). Any temporal restriction on causation denying that an effect may precede one of its causes would be seriously at odds with this important aspect of

theoretical physics.

4.4 Deterministic Causation: Focus on Pre-Emption

Many anti-reductionists have cited cases of causation under determinism that make trouble for Humean analyses. Among these, epiphenomena cases have played a central role. The basic problem all of these cases present is that analyses are prone to count as cause and effect two events that, in fact, only share a common cause (see Scriven 1966: 259 and Carroll 1994: 127–34). In recent discussions of deterministic causation, however, pre-emption cases have garnered tremendous attention. It is a good time to advance a new argument for anti-reductionism that builds on a preemption case.

The starting point will be one of Schaffer's (2000*b*) cases of *trumping preemption*. It is a law of magic that the first spell cast on a given day will match the enchantment that midnight. Suppose that at noon Merlin casts a spell (the first of the day) to turn the prince into a frog, that at 6:00 p.m. Morgana casts a spell (the only other that day) to turn the prince into a frog, and at midnight the prince becomes a frog (Schaffer 2000*b*: 165; also see McDermott 1995: 530; Menzies 1996: 95; Ehring 1997: 21–31).

A simple counterfactual analysis holding that *c* caused *e* if and only if *e* wouldn't have occurred if *c* hadn't occurred has the mistaken consequence that Merlin's casting the spell did not turn the prince into a frog. The standard way for counterfactual theories to try to sidestep problems with pre-emption—an appeal to the intermediate chain of events—does not help with this case. The sticking point is that there is no intermediate event between Merlin's spell and the enchantment—the spell acts directly.[1] The problems presented by pre-emption cases extend in a straightforward way to many other kinds of reductive analysis. These kinds of cases are the basis for Ehring's (1997: 18–49) rejection of a wide range of reductive analyses of causation.

Schaffer's case appears to leave a little room for the reductionist to manoeuvre. There are differences in the non-causal facts that seem to determine that it is Merlin's spell that is the cause. The seemingly pertinent non-causal facts are that Merlin's spell was the first spell cast that day, Morgana's spell was not the first cast that day, and that it is a law that the first spell cast on any given day matches the enchantment that midnight. But do these differences really determine the causal facts? That is not clear. There is nothing causal about Schaffer's law of magic; that law does not say that the first spell cast on a given day *causes* the enchantment at midnight—it only says that the first spell *matches* the enchantment at midnight. This observation is no objection to Schaffer's example. As Schaffer tells the story, we naturally assume that it is a causal law that (only) the first spell cast on a given day causes the enchantment at midnight. Nothing seems problematic about that natural assumption. So, we correctly conclude that Merlin turned the prince into a frog. The point of the observation is that philosophers have failed to recognize just how powerful the example is. Yes, it could be a law that only the first spell cast on a given day causes the enchantment at midnight. That is consistent and a natural thing to assume given Schaffer's description of the example. But, it is also consistent with Schaffer's description that instead it be a causal law that only the last spell cast causes the enchantment at midnight. Nature might work on a principle of least

effort: why use the spell requiring the action from the greater temporal distance when a closer one is available? If this were the case, then what we should conclude is that it was Morgana's spell, not Merlin's, that turned the prince into a frog. There is nothing more suspicious about this alternative way of filling in the details than the way that Schaffer actually had in mind.

The difference between the possible world in which Morgana's spell did the trick and the one in which Merlin's spell did the trick is a difference in the *causal* laws. What we have here is our first underdetermination example, two possible worlds that agree on their non-causal facts but disagree about what causes what. Given only the non-causal facts of Schaffer's case, it surely could be that Merlin did the trick. But, if one is prepared to accept this judgement, it seems that one should also be prepared to accept that, given only the non-causal facts of Schaffer's original case, it could also be that it was Morgana who did the trick. The anti-reductionist will be happy to employ trumping-style pre-emption cases as Schaffer, Ehring, and others do against counterfactual and other reductive analyses of causation. But the anti-reductionist should take the next step by concluding that causation does not supervene on the non-causal facts.

4.5 Indeterministic Causation

Schaffer (2000*a*: 40) has another useful magical example, one involving chance: 'Imagine that Merlin casts a spell with a .5 chance of turning the king and prince into frogs, that Morgana casts a spell with a (probabilistically independent) .5 chance of turning the prince and queen into frogs, and that the king and prince, but not the queen, then turn into frogs.' This is labelled a case of overlapping because the effects intended by Morgana and Merlin overlap —the sorceress and the sorcerer are both trying to turn the prince into a frog. The overlap is partial, though. Through her single spell, Morgana also wants to turn the queen into a frog; while through his single spell, Merlin also means to turn the king into a frog. Again it is assumed that, when they work, spells work directly, not through any intermediate events.

The causal facts about this case seem to be pretty straightforward. Since it was the king and the prince, and not the queen and prince, that became amphibians, it was Merlin's spell that was effective; Merlin, not Morgana, caused the prince to be a frog. But, these facts cut to the heart of the standard ways of dealing with probabilistic causation. If we consider a simple conditional-probabilities account, we will pick up on the fact that Morgana's spell raises the probability that the prince meets the amphibious fate. If we consider a simple counterfactual probabilistic account, what will be relevant is the fact that, if Morgana had not cast her spell, then the chance that the prince would become a frog would have been significantly less than it actually was. These analyses get the case wrong; they say Morgana's spell was effective.

Matters can be taken one step further. Suppose Merlin and Morgana both cast spells with a fifty–fifty chance of turning the prince into a frog. Neither is concerned with anyone else; they are both just after the prince. Like the previous case, this case involves overlapping; it is just that now the overlap is complete. What happens is that the prince turns into a frog (Schaffer 2000*a*: 45). Did Morgana turn the prince into a frog? Or was it Merlin? There seem to be at least two possibilities. The first is that Merlin did and that Morgana did not. The second is that Morgana did and Merlin did not. Nothing non-causal about the situation determines which is

the case. The case is structurally similar to cases endorsed by Scriven (1971: 62–4), Mackie (1974: 42–3), Foster (1979: 169–70), Armstrong (1983: 133; 1997: 203), Tooley (1984: 108–10; 1987: 199–202; 1990*a*: 274–8; 1990*b*: 225–8), Woodward (1990*b*: 214–16), and Carroll (1994: 137–8).

One more, devilishly simple, underdetermination case: suppose that there is an event, *k*, that immediately follows an event, *j*. Events such as *k* have a certain small probability of occurring at any time or place. There need not be any event or set of conditions that precedes the occurrence of this type of event; sometimes they just happen. *k*-type events do, however, tend to pop up more often just after a *j*-type event occurs. Remember there are lots and lots of *k*-type events that occur nowhere near any *j*-type events. Indeed, most of the *k*-type events occur without any connection to a *j*-type event. Still, in a high percentage of the cases, when a *j*-type event occurs a *k*-type event occurs immediately thereafter. (See Tooley 1990*a*: 278–9; 1990*b*: 229–30; 2003: 401–3; Woodward 1990*b*: 217; Carroll 1994: 140; and Armstrong 1999: 178–9; [2001] 2004: 449–50.) As already noted, what happens in our case is that *j* occurs and *k* occurs immediately thereafter. Question: Did *j* cause *k*? There are two possibilities. The first is that *j* did cause *k* and the second is that *j* had nothing to do with *k*—that *k* occurred uncaused as *k*-type events often do.

These cases are a serious challenge to the possibility of analysing causation. In the partial overlap case, there was the fact that the king turned into a frog that made it clear that it was Merlin's spell, not Morgana's, that was effective. The presence of that fact gives some hope to those who want to reduce causation. There is at least a symptom indicating that there might be some underlying truth-maker for the causal facts. The complete overlap case offers no hope at all, and neither does the subsequent case with an event that had a chance of being uncaused. With the sorcerers, there are no non-causal facts that determine whether Merlin was the cause. With *j* and *k*, there are none that determine whether *j* caused *k*.

5. REACTIONS TO THE ARGUMENTS

5.1 Anti-Reductionism is Uninformative

Ehring (1997: 62) has this to say about causation and non-supervenience:

One positive reason to reject, or at least not to quickly embrace, nonsupervenience is its relative lack of informativeness. Reductionist programs, if successful, are more philosophically enlightening. With reductionist accounts, we gain philosophical understanding into the nature of causation and its link with other important aspects of the world. Nonsupervenience cannot offer this. Hence, unless there are fairly strong arguments for nonsupervenience, we ought to pursue a reductionist program.

The point of analysis is to provide illumination about causation. Ehring finds anti-reductionism disappointing in what it can do toward attaining this goal.

This concern, however, is mistaken in presupposing that the ways in which anti-reductionism is uninformative do anything to establish a presumption in favour of reductionism. The anti-reductionist denies that there is any interesting answer to the question of how causes bring about their effects in suitably basic interactions (see Broad 1925: 453–4). But failure to answer that question cannot count against the anti-reductionist and in favour of the reductionist without begging the question. As the anti-reductionist sees it, the reductionist commits as serious a transgression by providing an answer to a question that does not have one. Furthermore, this reaction to the anti-reductionist's arguments underestimates the explanatory value of anti-reductionism. Indeed it does so in two ways. First, establishing anti-reductionism is itself to reveal something informative about causation. Second, there are analyses available to anti-reductionists that provide additional illumination. Analyses of causation in causal terms need not be trivial; they can make substantive and informative claims (see Woodward 2003: 20–2).

The bit of truth in this concern about uninformativeness rests in the fact that, unlike many reductionists, anti-reductionists have no built-in story to tell about certain paradoxical or puzzling features of causation. As will be discussed in sect. 6, there are some puzzling issues about causation that really do cry out for attention. For example, there is the issue of whether causation is transitive. Paradoxically, there is reason to think it is transitive and reason to think it is not. Often, by endorsing a particular analysis, philosophers commit themselves to a position on the transitivity of the causal relation or on another puzzling issue. Since the thesis of anti-reductionism itself does not commit one even indirectly to any position about what the causal relation is like, that thesis is silent on the puzzling issues. Especially in so far as the issues generate paradoxes, something more needs to be said about them by the anti-reductionist, especially if the anti-reductionist hopes to remain a realist about causation.

5.2 Anti-Reductionism Courts Scepticism

This concern is that we must lack causal knowledge that we usually presume ourselves to have if there is no non-causal analysis of causation. The idea is that, since causation is not directly experienced, without a non-causal analysis there is no way we could have that knowledge. The lack of an analysis, or worse the failure of causation to supervene on the non-causal, blocks any inferential path to causal knowledge.

Though a full response to this concern is not possible here, it is clear that anti-reductionism is not at stake. The demand for either direct access to causation or an analysis in experiential terms is a dangerous one. Berkeley's route to idealism about material objects, a view few take seriously, is the prime example of the trouble such reasoning brings. In this regard, it is helpful to keep in mind the similarities between our knowledge of material objects and our knowledge of causation. It is plausible that certain causal facts and certain facts with causal commitments are directly observable, at least in the weak sense described in sect. 2.4. In this, causal facts are no different from facts about material objects and events. And, of course, this way of attaining causal knowledge is open to sceptical attacks. But, in this regard, causal knowledge is no different from any of our other knowledge. There are sceptical arguments that seem to show we do not know much of *anything*; evil demon and other relevant-alternative

arguments are the most frustrating of the bunch. As important as sceptical reasoning is for philosophical investigation, it is doubtful that it could have any distinctive consequences for causation, since it is as compelling about material objects, events, and many concepts as it is about causation.

5.3 The Crucial Intuitions are Feeble and Foggy

Reductionists will acknowledge that the argument from sparseness (sect. 4.2) and the argument from directionality (sect. 4.3) are serious issues, though they still believe that a reductive analysis is possible—they insist that the non-causal base is rich enough to account for the directionality and the other features of causation. They can take this stance because the sparseness and directionality arguments do not deductively establish the anti-reductive conclusion.

The underdetermination examples from sects. 4.4 and 4.5 are a different matter. If the possible worlds described really are possible, then anti-reductionism follows validly. What have reductionists had to say about these kinds of examples? Not as much as one might expect, but they have found the intuitions behind the cases to be sufficiently weak to warrant taking the issue to be a 'don't-care question' (Woodward 2003: 383n.):

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. (ibid. 85)[2]

Nevertheless, the don't-care attitude is not warranted. The pre-theoretical intuitions are decidedly strong and clear. What is unclear is how the pre-theoretical intuitions are to be accommodated theoretically within favoured analyses, and it is this that leads reductionists to contend that the intuitions are inconsequential. At the very least, the anti-reductionist is owed an account of why the intuitions arise if they are not accurate.

Schaffer has recently taken the cases seriously. That is not surprising given the similarity of the underdetermination examples to his own partial overlap and trumping cases, and his desire to maintain supervenience. Schaffer recognizes the intuitions, but reports having strong countervailing intuitions that causal facts cannot float free. He also says what he thinks is really going on causally in a minor variation of the complete overlap case, one where it is clear that Merlin and Morgana did not both transform the prince: 'In such a case, I would answer that one of the spells caused the prince to transform, though it is ontologically indeterminate as to which. In some cases, there simply is no fact of the matter' (Schaffer 2008: 102n; see also Hitchcock 2004: 406–8). Schaffer (2008: 89) then offers an explanation of the intuitions that the anti-reductionist wants to take at face-value:

It seems to me that the reductionist can explain away the primitivist intuitions, from the

conceptual error of *reification.* Reification occurs when a mere concept is mistaken for a thing. We seem generally prone to this error. Our causal vocabulary allows us ... different descriptions, and this leaves us prone to positing ... different possibilities.

We can *say* that Merlin caused the transformation and Morgana did not. We can *say* that Morgana caused the transformation and Merlin did not. But, Schaffer claims, our ability to report (and conceptualize) in this way mistakenly leads us to think that there are genuine possibilities corresponding to those differences, that there could be something in nature that makes exactly one of these two conjunctive causal statements true.

Despite what Schaffer says, it is the can't-float-free intuition that has an easy explanation. It stems from the ubiquity of the reductive stance throughout contemporary analytic philosophy (and a possible reification of causation as an entity—see sect. 5.4). Also, Schaffer's intuition is what we might call a theoretical or higher-order intuition; it is not as straightforward as an intuition about the application of an ordinary concept to a hypothetical case. Other things being equal, we should trust the latter sort of judgements much more than we do the former. Furthermore, the anti-reductionist is likely to have his or her own higher-order intuitions about the applicability of our concepts—that a world could not be such that one of the two spells caused the transformation though it is not true that the first one did and it is also not true that the second one did.

5.4 Anti-Reductionism is Ontologically Extravagant

There is something else behind Schaffer's can't-float-free intuition and the subsequent conflict with the anti-reductionist intuitions. He says, 'even if there is some residual intuitiveness to the argument from causal differences, surely it is not sufficiently powerful to overturn the push for ... economical theory. After all, such a highly questionable intuition hardly seems sufficient to generate the sort of necessity needed to blunt Occam's Razor' (Schaffer 2008: 91). Lewis (1986: 180) says something similar about an overlap case and 'a metaphysical burden quite out of proportion to its intuitive appeal'. The worry is that if there really is something about the world that determines whether it was Merlin or Morgana that caused the transformation, then the world has some mighty funny things in it: hidden features that we otherwise had no reason to posit. The idea is that we can avoid the ontological commitment merely by ignoring the intuitive judgements and adopting a different stance on the underdetermination cases. Here it is ontological concern that drives the reductionist.

The anti-reductionist, however, is not committed to a mysterious ontology. The ontology of causation is independent of the issue of the supervenience and analysability of causation. Reification is a reductionist mistake. That the eruption of Mt Vesuvius caused the destruction of Pompeii, that Marvin hit Tommy, and that Merlin turned the prince into a frog, prima facie, bring no ontological commitment to anything beyond the eruption, the destruction, Marvin, Tommy, Merlin, the prince, and a frog. All these facts commit us to are events and objects, not anything anyone should be worried about. Even truths expressed by fact-causation sentences and states-of-affairs-causation sentences present no problem. These sentences can be rendered

ontologically innocuous: 'That Merlin cast his spell caused the prince to turn into a frog' only says 'The prince turned into a frog because Merlin cast his spell.' 'Vesuvius's erupting caused Pompeii's being destroyed' only says 'Pompeii was destroyed because Vesuvius erupted.' So, prima facie, ontologically speaking, we are only committed to Merlin, the prince, a frog, a spell, Vesuvius, and Pompeii. An austere anti-reductionism is not a non-starter.[3] More importantly, even if it were, there is certainly nothing about the anti-reductionist's intuitive judgements *regarding the underdetermination examples* that is ontologically reckless. Any arguments showing that causal facts commit us to something over and above objects and events would surely in the first instance show that we are committed to the added ontology by the more ordinary cases of causation just mentioned. How could Merlin's causing the prince to turn into a frog commit us ontologically to something besides Merlin, the prince, and a frog, and Vesuvius's causing Pompeii to be destroyed not commit us to something besides Vesuvius and Pompeii?

6. CONCLUSION

There have emerged in the recent literature numerous interesting issues about causation that transcend reductionism versus anti-reductionism.

The best example of such an issue is the matter of the transitivity of causation. Is causation transitive? It is natural to assume that it is: causation is making-happen. How can an event make another event happen and that second event make a third event happen and it not be that the first also made the third happen? Isn't it bound to be true that the first event made the third event happen by making the second event happen? But there are also examples that suggest the opposite. Suppose Sally places a bomb outside Ralph's door and lights the fuse. Once Sally leaves, Melissa happens to arrive at Ralph's place. Seeing the bomb and being a friend of Ralph's, she defuses the bomb, rendering it harmless. It seems that Sally's placing the bomb in front of Ralph's door caused Melissa to defuse it. It also seems that Melissa's defusing the bomb caused Ralph not to be killed. But, it seems false that Sally's placing the bomb caused Ralph not to be killed. Transitivity is hardly the only example. Other issues that are transcendent in the same way include overdetermination, the difference between causes and conditions, the efficacy of omissions, and also the features of the language we use to express causal truths.

The prominence of these transcendent issues bodes well. More and more, philosophers are not digging in their heels defending their favourite reductive analysis, holding whatever convenient position will facilitate their defence. Rather they are, somewhat independently of any specific analysis, revisiting these fundamental issues in an open-minded and provocative manner. The questions are not: What is wrong with this analysis? Is there any way of revising the analysis to avoid the problem? Instead the questions are: Is causation transitive? What causes what in cases of overdetermination? Are conditions causes? Do omissions have effects? Is there something about how the verb 'to cause' works in our language that sheds light on these puzzling issues? The preceding questions are all engaging and important. As was mentioned in sect. 5.1, since the thesis of anti-reductionism itself does not commit one to any particular views on these matters, and especially since some of the issues may generate

paradoxes, these questions may be crucial for the anti-reductionist. Fortunately, the anti-reductionist is also in a good position to address them, not having biases stemming from some favoured (and evidently false!) non-causal analysis of causation.

FURTHER READING

The place to begin a study of anti-reductionism is with Woodward 1990*b*. Additional arguments in support of the view can be found in Carroll 1994. Though they include sophisticated sorts of reductions of causation, the views of Armstrong and Tooley were central to the most recent development of anti-reductionism. Armstrong [2001] 2004 and Tooley 1990*b* are nicely refined presentations of their views. Scriven 1971 is an underappreciated gem of a paper. Schaffer (2008) provides the most interesting critical discussion to date of the underdetermination examples.

REFERENCES

ANSCOMBE, G. E. M. ([1971] 1981). *Causality and Determination: An Inaugural Lecture*. Cambridge: Cambridge University Press; repr. in her *The Collected Papers of G. E. M. Anscombe,* ii. *Metaphysics and the Philosophy of Mind*. Minneapolis: University of Minnesota Press, 133–47.

ARMSTRONG, D.M. (1983). *What is a Law of Nature?* Cambridge: Cambridge University Press.

—— (1993). 'Causes are Perceived and Introspected', *Behavioral and Brain Sciences* 16: 29.

—— (1997). *A World of States of Affairs*. New York: Cambridge University Press.

—— (1999). 'The Open Door: Counterfactual vs. Singularist Theories of Causation', in H. Sankey (ed.), *Causation and Laws of Nature*. Dordrecht: Kluwer, 175–85.

—— ([2001] 2004). 'Going Through the Open Door Again: Counterfactual vs. Singularist Theories of Causation', in G. Preyer and F. Siebelt (eds.), *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*. New York: Rowman & Littlefield, 163–76; repr. in J. Collins, E. J. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 445–57.

BLACKBURN, S. (1990). 'Hume and Thick Connexions', *Philosophy and Phenomenological Research* 50 suppl.: 237–50.

BRAND, M. (1975). 'On Philosophical Definitions', *Philosophy and Phenomenological Research* 36: 151–72.

—— (1976). 'Introduction: Defining Causes', in M. Brand (ed.), *The Nature of Causation*. Urbana: University of Illinois Press, 1–44.

BROAD, C. (1925). *The Mind and its Place in Nature*. London: Routledge & Kegan Paul.

CARROLL, J. (1991). 'Property-Level Causation?' *Philosophical Studies* 63: 245–70.

—— (1994). *Laws of Nature*. New York: Cambridge University Press.

—— (2008). 'Nailed to Hume's Cross?', in T. Sider, D. Zimmerman, and J. Hawthorne (eds.), *Contemporary Debates in Metaphysics*. Oxford: Blackwell, 67–81.

CARTWRIGHT, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Clarendon.

DEVITT, M. (1980). '"Ostrich Nominalism" or "Mirage Realism"?' *Pacific Philosophical Quarterly* 61: 433–9.

DUCASSE, C. ([1926] 1993). 'On the Nature and Observability of the Causal Relation', *Journal of Philosophy* 23: 57–68; repr. in E. Sosa and M. Tooley (eds.), *Causation*. Oxford: Oxford University Press, 125–36.

EHRING, D. (1997). *Causation and Persistence*. New York: Oxford University Press.

ELLIS, B. (2002). *The Philosophy of Nature: A Guide to the New Essentialism*. Montreal: McGill-Queen's University Press.

FOSTER, J. (1979). 'In *Self*-Defense', in G. Macdonald (ed.), *Perception and Identity*. Ithaca, NY: Cornell University Press, 161–85.

GOTT, R. (2001). *Time Travel in Einstein's Universe: The Physical Possibilities of Travel Through Time*. Boston: Houghton Mifflin.

GOTSHALK, D. (1930). 'Causality as an Ontological Relation'. *Monist* 40: 231–55.

—— (1931). 'Of the Nature and Definition of a Cause', *Philosophical Review* 40: 469–77.

HITCHCOCK, C. (2004). 'Do All and Only Causes Raise the Probabilities of Effects?', in J. Collins, E. J. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge: MIT, 403–17.

LAMPRECHT, S. (1929). 'Causality', in *Essays in Honor of John Dewey*. New York: Henry Holt and Company, 191–205.

—— (1930). 'Of a Curious Reluctance to Recognize Causal Efficacy', *Philosophical Review* 39: 403–14.

LEWIS, D. K. ([1973] 1986). 'Causation', *Journal of Philosophy* 70: 556–7; repr. in his *Philosophical Papers II*. Oxford: Oxford University Press, 159–72.

—— (1986). 'Postcripts to "Causation"', in his *Philosophical Papers II*. Oxford: Oxford University Press, 172–213.

MCDERMOTT, M. (1995). 'Redundant Causation', *British Journal for the Philosophy of Science* 46: 523–44.

MACKIE, J. (1974). *The Cement of the Universe*. Oxford: Clarendon.

MENZIES, P. (1996). 'Probabilistic Causation and the Pre-Emption Problem', *Mind* 105: 85–117.

—— (1999). 'Intrinsic versus Extrinsic Conceptions of Causation', in H. Sankey (ed.), *Causation and Laws of Nature*. Dordrecht: Kluwer, 313–29.

—— and PRICE, H. (1993). 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science* 44: 187–203.

PETERSON, J. (1898). 'The Empirical Theory of Causation'. *Philosophical Review* 7: 43–61.

PRICE, H. (2007). 'Causal Perspectivalism', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press, 250–92.

QUINE, W. (1948). 'On What There Is', *Review of Metaphysics* 2: 21–38.

RUSSELL, B. (1912–13). 'On the Notion of Cause', *Proceedings of the Aristotelian Society* 13: 1–26.

SCHAFFER, J. (2000*a*). 'Overlappings: Probability-Raising without Causation',

Australasian Journal of Philosophy 78: 40–6.

—— (2000*b*). 'Trumping Preemption', *Journal of Philosophy* 97: 165–81.

—— (2008). 'Causation and Laws of Nature: Reductionism', in T. Sider, D. Zimmerman, and J. Hawthorne (eds.), *Contemporary Debates in Metaphysics*. Oxford: Blackwell, 82–107.

SCRIVEN, M. (1966). 'Causes, Connections, and Conditions in History', in W. Dray (ed.), *Philosophical Analysis and History*. New York: Harper & Row, 238–64.

—— (1971). 'The Logic of Cause', *Theory and Decision* 2: 49–66.

—— (1975). 'Causation as Explanation', *Noûs* 9: 3–16.

TAYLOR, R. (1966). *Action and Purpose*. Englewood Cliffs, NJ: Prentice-Hall.

TOOLEY, M. (1984). 'Laws and Causal Relations', in P. French, T. Uehling, and H. Wettstein (eds.), *Midwest Studies in Philosophy*. Minneapolis: University of Minnesota Press, ix. 93–112.

—— (1987). *Causation: A Realist Approach*. Oxford: Oxford University Press.

—— (1990*a*). 'The Nature of Causation: A Singularist Account', in D. Copp (ed.), *Canadian Philosophers, Canadian Journal of Philosophy* suppl. 16: 271–322.

—— (1990*b*). 'Causation: Reductionism versus Realism', *Philosophy and Phenomenological Research* 50: 215–36.

—— (2003). 'Causation and Supervenience', in M. Loux and D. Zimmerman (eds.), *Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 386–434.

VON WRIGHT, G. (1971). *Explanation and Understanding*. Ithaca, NY: Cornell University Press.

—— ([1973] 1993). 'On the Logic and Epistemology of the Causal Relation', in P. Suppes (ed.), *Logic, Methodology, and the Philosophy of Science*. Amsterdam: North-Holland, iv. 293–312; repr. in E. Sosa and M. Tooley (eds.), *Causation*. Oxford: Oxford University Press, 105–24.

—— (1974). *Causality and Determinism*. New York: Columbia University Press.

WOODWARD, J. (1990*a*). 'Laws and Causes', *British Journal for the Philosophy of Science* 41: 553–73.

—— (1990*b*). 'Supervenience and Singular Causal Statements', in D. Knowles (ed.), *Explanation and Its Limits*. Cambridge: Cambridge University Press, 211–46.

—— (1994). Review of P. Humphreys, *The Chances of Explanation. British Journal for the Philosophy of Science* 45: 353–74.

—— (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

CHAPTER 14
CAUSAL MODELLING

CHRISTOPHER HITCHCOCK

1. INTRODUCTION

'Causal modelling' is a general term that applies to a wide variety of formal methods for representing, and facilitating inferences about, causal relationships. The end of the twentieth century saw an explosion of work on causal modelling, with contributions from such fields as statistics, computer science, and philosophy; as well as from more subject-specific disciplines such as econometrics and epidemiology. In this entry I will focus on two programmes that have attracted considerable philosophical attention, one due to the computer scientist Judea Pearl and his collaborators, and the other to the philosophers Peter Spirtes, Clark Glymour, and Richard Scheines.

Unlike the more traditional philosophical accounts of causation canvassed in Part II of this volume, the causal modelling programmes of Pearl and Spirtes, Glymour, and Scheines do not attempt to analyse causation in terms of anything else. Nonetheless, they do establish interconnections between causal relationships on the one hand, and regularities, counterfactuals, interventions, and probabilities on the other; hence the causal modelling programmes make contact with more traditional programmes at a number of points.

While the most common use of causal models is to facilitate causal inference, this application will not be the focus of this chapter. Causal inference is discussed in detail in Ch. 23. Instead, this chapter will offer a much simplified presentation of causal models that emphizes various points of philosophical interest.

2. ILLUSTRATION

We begin with an illustration of a deterministic causal model, based on one presented in Pearl (2000: ch. 3). Suppose that in a certain agricultural region, oat crops are threatened by eelworms. We are interested in whether the use of a certain fumigant is effective in protecting oat yields, and to what degree. We can't simply compare the oat yields of those who use the fumigants to those who do not, for it may be that only the farmers who are suffering the worst infestations choose to use the fumigants. There are other complications as well: it is possible that the fumigant directly affects the oat crop; and the population of eelworms is independently regulated by birds that prey on them.

A deterministic causal model is an ordered pair , where V is a set of variables, and E is a set

of equations relating the values of those variables. Our model might include the variables:

E_1 the population of eelworms before the time at which the fumigant is (or would be) applied

F the quantity of fumigant used (possibly zero)

B the population of birds that prey on eelworms

E_2 the population of eelworms after the time at which the fumigant was (or would have been) applied

Y the yield of the oat crop.

Note that these are all quantitative variables, rather than properties or events. Philosophers have traditionally taken the latter (or sometimes related entities such as facts or tropes) to be the relata of causation (see Ch. 19 of this volume). In scientific contexts, however, it is more common to describe causal relationships in terms of variables.

The relationship between these variables is captured in a set of equations:

$$E_1 = e_1$$
$$F = f_F (E_1)$$
$$B = f_B (E_1)$$
$$E_2 = f_E (E_1, F, B)$$
$$Y = f_Y (F, E_2)$$
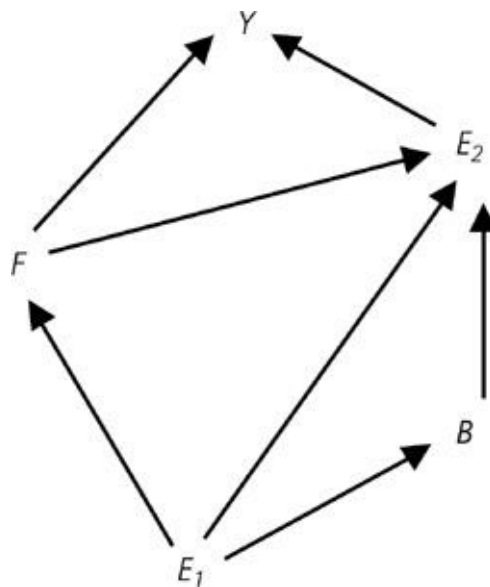
Fig. 14.1

These equations are called *structural equations*, since they are intended to describe the underlying causal structure, and not merely to describe regularities among the values of the variables. The full significance of this distinction will become apparent in sect. 4 below. Pearl (2000) says that these equations describe *mechanisms*, although this usage seems to differ from that in recent philosophical discussions of mechanisms (see Ch. 15).

Note that there is one equation for each variable. Since the value of E_1 is not affected by the value of any of the other variables, its value is given from outside the system. This variable is *exogenous*.[1] All the other variables are endogenous. In our example, each variable is introduced on the left-hand side of a structural equation before appearing on the right. A set of equations that has this feature (or that can be reordered so as to have this feature) is *acyclic*. We will consider only acyclic structures here. If one variable appears on the right-hand side of an equation, it is said to be a *parent* of the variable appearing on the left. Hence E_1 is a parent of $F, B,$ and E_2; F is a parent of E_2 and Y, and so on. The terms *child, ancestor,* and *descendant* may be defined from 'parent' in the obvious way.

The qualitative relationships among the variables may be captured using a *directed graph*. Each node in the graph corresponds to a variable in the model, and an arrow is drawn from one variable to another just in case the former is a parent of the latter. A series of arrows aligned tip to tail constitutes a *path* through the graph. A graph is *acyclic* (representing an acyclic set of equations) just in case no path closes back on itself to form a loop. The graph corresponding to our model is shown in .

3. Regularities and Predictions

A causal model is a mini-theory that entails certain regularities and allows us to make predictions on the basis of given observations. Let's suppose that we observe the bird population to be b. How may we use the model to predict the values of the other variables? We do this by substituting the value b for the variable B in all equations, and then solving for the resulting system of equations. This may or may not yield a definite value for the other variables. For example, if the function f_B is invertible, the model predicts that when B takes the value b, then F takes the value $\left(f_B^{-1}(b)\right)$. Thus the model may entail a number of regularities beyond those directly represented in the structural equations.

While the equations of a model entail regularities, they fall well short of the status of *laws*. Consider the equation $Y = f_Y(F, E_2)$. It is wildly implausible that the quantity of fumigant used and the population of eelworms are the only factors that influence the size of an oat yield. Other factors will surely include the amount of sunshine and rainfall, average temperatures, the composition of the soil, the number of seeds planted, etc. The equation $Y = f_Y(F, E_2)$ will only hold for some values of these variables. In order for the model to apply to multiple oat fields, or to the same oat field in different growing seasons, these additional factors must remain stable. Thus our causal model does not entail that *whenever* the quantity of fumigant used is f, and the eelworm population e, the oat yield will be $f_Y(f, e)$. The causal model by

itself does not tell us whether it is meant to apply to a single oat crop, to multiple fields, or to multiple growing seasons; it is up to the user of the model to choose whether and when to rely on the model to make predictions. If the predictions of the model are not borne out by a particular oat crop, that shows that the model does not correctly describe the causal relations among the relevant variables within that crop. In this way, causal models may be tested empirically.

4. INTERVENTIONS AND COUNTERFACTUALS

A causal model can also be used to predict the effects of certain interventions, and to evaluate the truth values of certain counterfactuals (non-backtracking counterfactuals, in the sense of Lewis ([1979] 1986); see Ch. 8 for further discussion). There is an important pragmatic difference between counterfactuals and interventions: we are typically interested in knowing the truth values of counterfactuals after the fact, whereas we are usually interested in evaluating the consequences of potential interventions before they are carried out. Apart from this difference, however, the mechanics of evaluating counterfactuals and interventions is the same. Suppose, for example, we wish to know the effects of an intervention that sets the level of fumigant used to f. To say that the fumigant level is set by an intervention is to say that the normal causal factors that influence the value of this variable are overridden: the level of fumigant used will be f regardless of what the original eelworm population is. We represent the effect of this intervention by *replacing* the equation for F with a new one that merely stipulates the value of F. We can then calculate the effects of this intervention by solving for the new set of equations. We evaluate the consequences of introducing a counterfactual antecedent $F = f$ in exactly the same way. The graph corresponding to the new system of equations will look like Fig. 14.1, except that the arrow into F will be removed.

This procedure only represents the effects of surgical or idealized interventions that interfere with only one of the mechanisms represented in the causal model. We could, in principle, affect the amount of fumigant that a farmer chooses to use by intervening on the original eelworm population, but this would not be a surgical intervention on the level of fumigant used, and the results of such an intervention would not be accurately predicted by replacing the equation for F. For detailed discussion, see Woodward (2003), and well as Ch. 11 above. According to Lewis ([1979] 1986), the antecedents of counterfactuals are to be thought of as coming about by 'small miracles'. These small miracles are just extreme cases of surgical interventions. We need not posit miracles, however: since the equations in the causal model do not represent laws, we do not need to suppose that the modification of a structural equation involves any kind of miracle.

The procedure for evaluating interventions and counterfactuals differs from the procedure for making predictions on the basis of observations. By replacing the equation for F, rather than substituting the value f for F, we get the new equation $F = f$, rather than the equation $f = f_F(E_1)$. In this way, the hypothetical change in the value of F does not backtrack and affect the value of E_1. Because of the way in which counterfactuals and interventions are handled, the form of the equations becomes important. The equation $F = f_F(E_1)$ is algebraically equivalent

to the equation $E_i = f_F^{-1}(F)$ (assuming f_F to be invertible), but the two equations entail different counterfactuals and intervention effects.

We are now better able to understand the central difference between the 'structural' equations, and equations such as $F = f_F\left(f_B^{-1}(B)\right)$ that merely describe regularities that are generated by the underlying causal structure. A structural equation remains unchanged when we intervene on the value of one of the variables on its right-hand side. For example, if we were to intervene on the bird population B, we would represent this by replacing the equation for B, but not any of the other structural equations. By contrast, if we were to intervene on the value of B, the equation $F = f_F\left(f_B^{-1}(B)\right)$ would cease to hold. This regularity is sustained only because of the way in which the value of B is actually brought about; if we intervene to set the value of B in some other manner the regularity breaks down. An equation that describes a genuine causal relationship, by contrast, will hold even when the values of the effect variables are determined in some new way. Since we are agents who frequently intervene in the natural order of things, it is easy to see why we would be particularly interested in those regularities that will continue to hold in the wake of our interventions. For further discussion of these issues, see Ch. 11.

Pearl (2000) very clearly considers the causal mechanisms represented by structural equations to be primitive, and takes the truth values of counterfactuals to be grounded in these mechanisms. Suppose, however, that there is some independent way of grounding the truth values of counterfactuals, perhaps in terms of the similarity metric among possible worlds described by Lewis ([1979] 1986). Then it will be possible to recover the system of structural equations among a set of variables from the set of true counterfactuals about values of these variables. In this way, causal models can become very useful tools for representing complex patterns of counterfactual dependence, and the causal modelling framework becomes a powerful extension of the counterfactual approach to causation (discussed in Ch. 8).

5. CAUSAL INTERPRETATION OF THE MODELS

What, exactly do causal models represent? In particular, what do the arrows in a graph such as Fig. 14.1 represent? Clearly, the arrows indicate that parent variables exercise some kind of causal influence over their children. Nonetheless, for a number of reasons it is awkward at best to say simply that an arrow from X to Y indicates that X causes Y.

First, an arrow does not tell us anything about the form of the quantitative relationship between two variables. For example, Fig. 14.1 contains an arrow from E_1 to E_2 and also an arrow from F to E_2. It is plausible that E_2 increases with increasing values of E_1 but decreases with increasing values of F. This information can be gleaned from the structural equations, but not from the graph alone. It would certainly be misleading, if not downright ungrammatical, to say that fumigant level causes later eelworm population. 'Affects', 'influences', or 'is causally relevant to' would be more appropriate terms here than 'causes'.

Second, there are causal relationships that are not marked with arrows. For example, the initial eelworm population will probably be causally relevant to the final oat yield, yet there is

no arrow directly connecting E_1 to Y. Why not? Our model tells us that if we intervene to fix the level of fumigant used, and also intervene to fix the later eelworm population, then a further intervention on the initial eelworm population will have no effect on the oat yield. If the initial eelworm population has an effect on the oat yield, this effect is mediated by fumigant level and the later eelworm population. The arrows thus represent only direct effects, causal influences that are unmediated by other variables in the variable set V.

As a result, a causal model will contain more information than could be conveyed by a simple listing of which variables causally influence which others. For example, if we were told that fumigant level affects both later eelworm population and crop yield, and that later eelworm population affects crop yield, this would not tell us whether fumigant level has a direct effect on oat yield, or only an indirect effect via its influence on the eelworm population. A causal model therefore tells us not only which variables have a causal influence on others, but it also tells us about the various pathways along which causal influence is exercised.

6. CAUSAL CONCEPTS

It is possible to define a number of different causal concepts from within the causal modelling framework. One we have already seen: X has a *direct effect* on Y, if and only if X is a parent of Y. Here are some other examples:

> X has a *total,* or *net, effect* on Y if and only if at least some interventions on the value of X yield different values of Y.
> The *causal effect* on Y of a change in the value of X from x to x' is the difference between the value of Y that would result from an intervention setting X to x', and the value of Y that would result from an intervention setting X to x. X has a *component,* or *path-specific* effect on Y along some specific path if and only if there is some set of values for the variables that do not lie on the path, such that when we intervene to set all those variables to those values, at least some interventions on the value of X will lead to different values of Y.
> X is *causally relevant* to Y, if and only if there is some set of variables, and some set of values of those variables, such that when we intervene to set all those variables to those values, at least some interventions on the value of X will lead to different values of Y.

The precise terminology and definitions differ from author to author, but these examples give some feel for the range of new concepts that may be introduced. In contrast to the way in which causation is often approached in philosophy, there is no presumption that any one of these concepts is uniquely deserving of the title 'causation'.

7. PROBABILITY

So far, we have considered only deterministic causal models, but it is often useful to employ causal models that make probabilistic, rather than deterministic predictions. A

probabilistic causal model is an ordered triple Pr>, where V is a set of variables as before, G is a directed graph over the variables in V representing the causal relations among them, and *Pr* is a probability distribution over the variables in V that represents empirical probability, as estimated by frequency data. Once again, we will consider only acyclic graphs. The probability distribution *Pr* can be tested directly against statistical data, but how can we test whether G accurately captures the causal structure that generates these statistical data? Presumably, the causal structure relating a set of variables will somehow constrain the probability distribution over the value of those variables. The most important such constraint is the *Causal Markov Condition.*[2]

8. THE CAUSAL MARKOV CONDITION

Let V be a set of variables, and let G be a directed acyclic graph representing the causal relationships among those variables. For any X in V, let PA(X) denote the set of parents of X in V, DE(X) be the set of descendants of X in V, and ND(X) the set of variables in V, other than X itself, that are not descendants of X.[3] A probability distribution *Pr* satisfies the *Causal Markov Condition* with graph G just in case:

For every X in V, and every set Y of variables in ND(X),
$$Pr(X|\ PA(X)\&Y) = Pr(X|PA(X)).$$
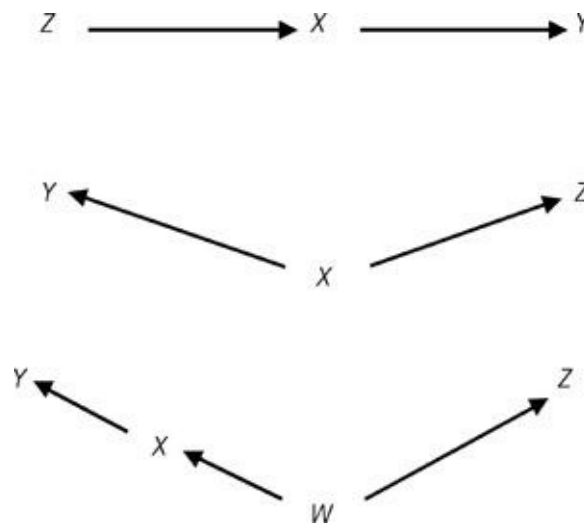
Fig. 14.2

This claim contains implicit universal quantifiers over the values of the variables.[4] A probabilistic causal model Pr> with an acyclic graph that satisfies the Causal Markov Condition is called a *causal Bayes net*.

The Causal Markov Condition says that a variable is conditionally independent of its non-descendants, given the values of its parents. In the terminology of Reichenbach (1956), every

variable is *screened off* from all its non-descendants by its parents. The Causal Markov Condition is a generalization of Reichenbach's *Common Cause Principle* (see Ch. 9); it entails, for example, that *X* will screen *Y* off from *Z* in each of the causal structures shown in Fig. 14.2. In general, a causal Bayes net will predict a number of conditional independence relations that can be tested against statistical data.

The Causal Markov Condition can be motivated by starting with a deterministic causal model, and introducing an element of uncertainty. Consider our causal model concerning oat crops. As we noted in sect. 3 above, it is very implausible that this model successfully incorporates all the factors that affect the values of the variables included in the model. We might reflect this by adding a number of additional exogenous variables to the model, representing the influence of unknown factors. Such variables are often referred to as *error variables*. We would then rewrite our equations thus:

$$E_1 = U_{E1}$$
$$F = f_F(E_1, U_F)$$
$$B = f_B(E_1, U_B)$$
$$E_2 = f_E(E_1, F, B, U_{E2})$$
$$Y = f_Y(F, E_2, U_Y)$$

An assignment of values to the error variables uniquely determines the values of all the variables in the model. In general, it will be impossible to know the values of the error variables, so we posit a probability distribution *Pr'* over the values of these variables. The probability distribution over the values of the error variables will induce a probability distribution *Pr* over the values of the other variables in the model. If the error terms are probabilistically independent; that is, if

$$Pr'(U_1 = u_1...U_n = u_n) = Pr'(U_1 = u_1) \times...\times Pr'(U_n = u_n)$$

for all u_i, $i = 1, ..., n$,[5] then the graph of Fig. 14.1 and the distribution *Pr* will satisfy the Causal Markov Condition (Pearl and Verma 1991).

The Causal Markov Condition will not hold for arbitrary sets of variables, however; it will typically fail if two variables in V share a common cause that is not itself included in V. It is plausible, however, that the condition will hold for variable sets that are *causally sufficient*, that contain all common causes of variables included in the set. Proponents believe that when causal models are chosen carefully, the Causal Markov Condition will hold frequently enough to make it a valuable tool in causal inference. See Ch. 23 for a full discussion of these issues.

9. THE MINIMALITY AND FAITHFULNESS CONDITIONS

Spirtes, Glymour, and Scheines (2000) introduce two other conditions to supplement the Causal Markov Conditions. Let G be a directed acyclic graph over V, and let *Pr* be a probability distribution over V that satisfies the Causal Markov Condition relative to G. The *Minimality Condition* says that no subgraph of G also satisfies the Causal Markov Condition.

The *Faithfulness Condition,* which is strictly stronger than the Minimality Condition, says that the probability distribution contains no conditional independence relations that are not entailed by the Causal Markov Condition. Either one of these conditions can fail if there is fortuitous cancellation of causal influences. Consider the causal graph in Fig. 14.1. According to this graph, fumigant levels have a direct effect on oat yields, and also an indirect effect via their effect on eelworm populations. Now suppose that these two effects always exactly cancel, so that there is no net effect of fumigant levels on oat yields. Then fumigant level and oat yield will be (unconditionally) independent. This independence is not implied by the Causal Markov Condition, and hence the Faithfulness Condition would be violated. The Minimality and Faithfulness Conditions function as simplicity assumptions that guide causal inference: if one discovers a probabilistic independence among variables, it is preferable to assume that this independence is the result of a simpler causal structure rather than a more complex causal structure with fortuitous cancellation of causal influences.

10. REDUCTION

Reichenbach (1956) hoped to reduce causal relations to probabilistic relations (where the probabilities were understood in terms of limiting relative frequencies). Cartwright ([1979] 1983) argued that this is not possible (see Ch. 9 above for discussion). The causal modelling framework provides a very powerful tool for exploring this issue. In asking whether it is possible to reduce causation to probabilities, one question we might ask is whether it is possible to offer an analysis along the lines '*C* causes *E* if and only if ... ' where the right-hand side makes reference only to probabilities, and not to causal relations. The causal modelling framework does not offer any kind of reduction along these lines. But there is a different question that we might ask. A probabilistic theory of causation such as that of Reichenbach (1956) or Cartwright ([1979] 1983) will impose constraints on the relationship between causal structure and probability relations, and we might ask whether the constraints imposed by the theory render the probability distribution over all the variables compatible with only one causal structure over those variables. If this is the case, then the probability distribution uniquely determines the causal structure, and we might reasonably be said to have provided a kind of reduction of causation to probability.

Let us say that two causal structures (represented by directed acyclic graphs) over a vertex set V are *statistically indistinguishable* if the probability distributions compatible with one are compatible with the other. We can define a number of different notions of statistical indistinguishability, depending upon whether we require that all, or only some, probability distributions compatible with the one to be compatible with the other, and depending upon which constraints we impose on the relation between causal structure and probability. Spirtes, Glymour, and Scheines (2000: ch. 4) explore the consequences of two different sets of conditions: the Causal Markov Condition plus the Minimality Condition, and the Causal Markov Condition plus the Faithfulness Condition. In general, it is not the case that a probability distribution will be compatible with only one causal structure, so prospects for a reduction of causation to probability do not look promising. Nonetheless, Spirtes, Glymour,

and Scheines do prove a number of interesting results; for example, any two causal structures over the same variable set may be embedded into larger structures that are statistically distinguishable.

11. EPISTEMOLOGY

The primary use of causal models is to facilitate causal inference. A detailed discussion of causal inference is presented in Ch. 23, but we may make some elementary observations here. We have seen how a causal model predicts certain regularities, probabilistic independences, and effects of interventions. These predictions may then be tested empirically, leading to the (partial) confirmation or disconfirmation of the causal model. In this regard, causal models function like other scientific theories; there is no *special* problem of assessing causal claims. This view is in stark contrast to the empiricist tradition in philosophy that has regarded causal claims as epistemically inaccessible and hence metaphysically suspect.

12. ACTUAL CAUSATION

We noted in sect. 6 that a number of different causal concepts can all be defined from within the causal modelling framework. Recently, there has been considerable interest in trying to offer a definition of *actual causation,* what is sometimes called *token* or *singular* causation, from within a causal modelling framework. The notion of actual causation plays a role in judgements of moral and legal responsibility, and for this reason it has been of considerable interest to philosophers and legal theorists. Indeed, much of the work in counterfactual approaches to causation, following the seminal work of Lewis ([1973] 1986) has been devoted to analysing actual causation—usually just called 'causation'—in counterfactual terms (see Ch. 8 for discussion). Given the strong connection between causal models and counterfactuals discussed in sect. 3, it is natural to think of causal modelling approaches as extensions of counterfactual approaches.

We may illustrate the problem of defining actual causation using a simple example of *pre-emption* (see Ch. 17 for discussion of different kinds of pre-emption). An assassin poisons his victim's drink; the victim drinks the poison and dies. If the assassin had not poisoned the drink, his back-up would have done so. The assassin poisoning the drink caused—was an actual cause of—the victim's death.

Let us represent this scenario using a deterministic causal model. The variable set V will be {A, B, D}, where the variables have the following interpretation:

$A = 1$ if the assassin poisons the drink, 0 otherwise;

$B = 1$ if the backup poisons the drink, 0 otherwise;

$D = 1$ if the victim dies, 0 otherwise:

Here the values of the variables represent the occurrence or non-occurrence of specific

events. The equations will be:

$$A = 1$$

$$B = 1 - A$$

$$D = \max \{A, B\}$$

The corresponding graph is shown in Fig. 14.3. These equations can be interpreted by means of the counterfactuals that they entail. The second equation entails that the back-up would have poisoned the coffee if the assassin had not; the third equation entails that the victim would not have died if neither the assassin nor the back-up had poisoned the coffee, and so on. From these equations, we can compute that $A = 1, B = 0$, and $D = 1$. In order to determine what would have happened if the assassin had not poisoned the coffee, we replace the first equation with $A = 0$. We can then compute that $B = 1$ and $D = 1$; our model correctly entails that the victim would still have died if the assassin had not poisoned the coffee.

The central idea behind the various attempts to define actual causation is that actual causation does not require counterfactual dependence within the original model, but only counterfactual dependence within some suitable modification of the original model. Suppose, for example, that we modify our model by replacing the second equation with $B = 0$, thus fixing B at its actual value of 0. In the resulting model, we still have $A = 1, B = 0$, and $D = 1$. But now changing A to 0 results in a change in the value of D, so D counterfactually depends upon A in this new model. The key question, then, is which modifications of the original model are permissible.
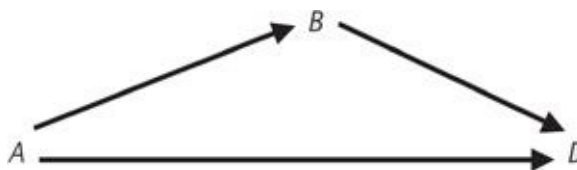
Fig. 14.3

Perhaps the simplest proposal is offered by Hitchcock (2001). In Fig. 14.3, we can see that there are two paths from A to D, one direct, and one indirect. We may think of the modification described in the previous paragraph as isolating the effect of A on D along the direct path. By fixing the variable B at its actual value, we prevent any causal influence from travelling along that path from A to D. Since D does depend upon A when B is thus frozen, A exerts a causal influence on D along the direct path. Hitchcock's proposal is that $X = x$ is an actual cause of $Y = y$ in a causal model if there is a path from X to Y that has the following property: when all the variables that lie off this path are held fixed at their actual values, Y

depends counterfactually upon *X*. Thus the modifications allowed by Hitchcock's proposal are those that fix all variables at their actual values, except for those that lie along some specific path.

While Hitchcock's proposal works well for cases of pre-emption of the sort just described, it has problems with cases of *overdetermination* (see Ch. 17). Suppose that we change the scenario so that both the assassin and the back-up poison the victim's drink. Again, we are inclined to say that the assassin's poisoning the drink is a cause of the victim's death, but there is no counterfactual dependence. Our model will be:

$$A = 1$$

$$B = 1$$

$$D = \max \{A, B\}$$

We can see that Hitchcock's proposal will fail: even if we hold *B* fixed at its actual value of 1, we do not have counterfactual dependence of *D* upon *A*.

The most powerful account of actual causation in the causal modelling framework is that of Halpern and Pearl (2001; 2005). Suppose we have a deterministic causal model , in which $X = x$ and $Y = y$. $X = x$ is an actual cause of $Y = y$[6] in this model if there exists a subset W = $\{W_i\}$ of V and a set $\{w_i\}$ of values of those variables (not necessarily their actual values in), such that: (1) when each W_i is set to w_i, *Y* counterfactually depends upon *X*; and (2) when each W_i is set to w_i, and any subset of variables in V\W are set to their actual values, then *Y* will still take its actual value of *y*. Clause (1) asserts the basic idea underlying causal modelling accounts of actual causation: that $X = x$ is an actual cause of $Y = y$ if *Y* counterfactually depends upon *X* in a suitably modified model. Clause (2) imposes a restriction on the modifications that can be made, in particular it requires that the new settings of the variables in W do not affect the value of the candidate effect variable *Y*.

This definition is complex, and readers are referred to the original papers for detailed discussion. We will here just make a few observations. First, according to the Halpern and Pearl definition, $X = x$ will count as a cause of $Y = y$ if *Y* counterfactually depends upon *X* in the original model; this just corresponds to the case where W is empty. Second, Halpern and Pearl's definition subsumes Hitchcock's as a special case. The set W will be the set of variables that lie off some path from *X* to *Y*, and the settings will be the actual values of these variables. Third, the Halpern–Pearl account can handle our example of overdetermination. Here we take W = $\{B\}$ and set *B* to 0. This setting is permissible under clause (2), since it does not change the value of *D*, regardless of whether we fix any other variables at their actual values. When we set *B* to 0, *D* counterfactually depends upon *A*, so $A = 1$ counts as an actual cause of $D = 1$.

Nonetheless, the Halpern–Pearl account does not always yield the intuitively correct

answer. Suppose, for example, that our assassin refrains from poisoning the drink, but that a bodyguard administers an antidote to the potential victim anyway.[7] Most people judge that the bodyguard's administration of the antidote was not an actual cause of the victim's survival—the antidote was in fact completely unnecessary given the lack of poison. But the Halpern–Pearl account treats this example in the same way that it treats overdetermination. Our model will be

$$A = 0$$

$$B = 1$$

$$D = \min \{A; 1 - B\}$$

$B = 1$ represents the bodyguard's administration of the antidote. Now we can choose W = {A}. Setting A to 1 does not affect the value of D, regardless of what else we hold fixed at its actual value, but when we set A to 1, D counterfactually depends upon B. Thus the Halpern–Pearl account rules that $B = 1$ is an actual cause of $D = 0$. One problem is that the Halpern–Pearl definition does not reflect the way in which the difference between actions and omissions affect our judgements of actual causation. Hitchcock (2007) offers an account that attempts to take this distinction into account. The analysis of actual causation remains an area of continuing research.

FURTHER READING

The two central works in the tradition surveyed here are Pearl (2000) and Spirtes, Glymour, and Scheines (2000). Both works are technical, but will reward a careful reading by anyone with a serious interest in causation. The last chapter of Pearl (2000) presents an informal and accessible overview of some issues in causal modelling, and Pearl (1999) presents an introduction to deterministic causal models and their connection with counterfactuals. Neapolitan (2003) is a good text book dealing with technical aspects of Bayes nets. Woodward (2003) contains an extended discussion of the relationship between causal models and interventions. The central papers on causal modelling approaches to actual causation are Halpern and Pearl (2001; 2005), and Hitchcock (2001).

REFERENCES

CARTWRIGHT, N. ([1979] 1983). 'Causal Laws and Effective Strategies'. *Noûs* 13: 419–37; repr.in N. Cartwright, *How the Laws of Physics Lie*. Oxford: Clarendon, 1983.
HALPERN, J., and PEARL. J. (2001). 'Causes and Explanations: A Structural-Model Approach—Part I: Causes'. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann, 194–202.

—— (2005). 'Causes and Explanations: A Structural-Model Approach—[Part I](): Causes' (expanded version). *British Journal for the Philosophy of Science* 56: 843–87.

HIDDLESTON, E. (2005). 'Causal Powers'. *British Journal for the Philosophy of Science* 56: 27–59.

HITCHCOCK, C. (2001). 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98: 273–99.

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

LEWIS, D. ([1973] 1986). 'Causation', *Journal of Philosophy* 70: 556–67; repr. in Lewis (1986), 159–72.

—— ([1979] 1986). 'Counterfactual Dependence and Time's Arrow', *Noûs* 13: 455–76; repr. in Lewis (1986: 32–52).

—— (1986). *Philosophical Papers II*. Oxford: Oxford University Press.

NEAPOLITAN, R. (2003). *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.

PEARL, J. (1999). 'Reasoning with Cause and Effect'. *Proceedings of the International Joint Conference on Artificial Intelligence*. San Francisco: Morgan Kaufmann, 1437–49.

—— (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

—— and VERMA, T. (1991). 'A Theory of Inferred Causation', in J. Allen, R. Fikes, and E. Sandewall (eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*. San Mateo, Calif.: Morgan Kaufman, 441–52.

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

SPIRTES, P., GLYMOUR, C., and SCHEINES, R. (2000). *Causation, Prediction, and Search*. 2nd edn. Cambridge, Mass.: MIT.

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

CHAPTER 15
MECHANISMS

STUART GLENNAN

1. INTRODUCTION

The term 'mechanism' is commonly used to refer to a variety of systems or processes that produce phenomena in virtue of the arrangement and interaction of a number of parts. The term was originally applied to products of human design like watches or water wheels, but has, at least since the seventeenth century been used equally to describe systems (like cells) or processes (like those that produce sunspots) that are of natural origin (Dijksterhuis 1969).

Mechanism is undoubtedly a causal concept, in the sense that ordinary definitions and philosophical analyses explicate the concept in terms of other causal concepts such as production and interaction. Given this fact, many philosophers have supposed that analyses of the concept of mechanism, while they might appeal to philosophical theories about the nature of causation, could do little to inform such theories. On the other hand, methods of causal inference and explanation appeal to mechanisms. Discovering a mechanism is the gold standard for establishing and explaining causal connections. This fact suggests that it might be possible to provide an analysis of causation that appeals to mechanisms.

There have been a variety of attempts to explicate the concept of mechanism and to deploy it in understanding causation and causal explanation. Salmon (1984) and Dowe (2000) treat mechanisms as a nexus of continuous physical processes. Bechtel and Richardson (1993), Glennan (1996), and Machamer, Darden, and Craver (2000) treat mechanisms as systems of interacting parts. As the process approach is discussed elsewhere (Ch. 10), I will focus on the mechanical systems approach, which for brevity I shall refer to simply as the mechanical approach.

Glennan, who is the most explicit in trying to develop a mechanical account of causation, characterizes mechanisms in this way: 'A mechanism underlying a behavior is a complex system which produces that behavior by the interaction of a number of parts according to direct causal laws' (Glennan 1996: 52). Glennan then suggests that two events are causally related when and only when they are connected by an intervening mechanism.

I will explore three issues surrounding the adequacy of the mechanical approach to causation. First, I consider whether the appeal to laws or invariant generalizations in characterizing interactions between parts of mechanisms either makes the mechanical theory circular or reduces it to a regularity, counterfactual, or manipulability theory. Second, I discuss Machamer, Darden, and Craver's argument that the proper understanding of the causal productivity of mechanisms requires the recognition of the novel metaphysical category of

activities. Third, I discuss the relationship between mechanical theories and process theories.

2. Laws, Generalizations, and Mechanisms

Machamer, Darden, and Craver (2000; hereafter MDC) and Woodward (2002) have raised concerns about Glennan's use of laws to characterize interactions between parts of mechanisms. According to one common understanding of scientific laws, laws must be exceptionless regularities of unrestricted scope. If one understands laws in this way, it is implausible to assume that all interactions between parts of mechanisms are law-governed, because the regularities involved in the operation of mechanisms can be fragile and subject to exceptions and breakdowns.

But this objection is largely an issue of terminology. Glennan explicitly adopts an alternative usage which understands laws to comprise a large class of non-accidental, counterfactual supporting generalizations, many of them of restricted scope.[1] For example, according to this usage Mendel's law of independent assortment really is a law, even though it is subject to exceptions (as when loci lie near to each other on the same chromosome). While this usage has the virtue of stressing that the honorific 'law' is applied to many generalizations of this character, Glennan has more recently (2002) borrowed terminology from Woodward (2000), describing such generalizations as direct invariant change-relating generalizations. These generalizations characterize interactions in which changes in the properties of one or more parts bring about changes in a property of another part. The requirement that these generalizations be direct dictates that the change of properties in one part brings about the change in the properties of another part without changing the properties of another part that is intermediate within a causal chain.[2]

Whatever one calls generalizations describing interactions between parts of mechanisms, critics can argue that that these generalizations do all the causal work. A mechanist must have some account of truth conditions for these generalizations, and the suspicion is that any real insight into the nature of causation will lie in these conditions rather than in the analysis of mechanisms. In particular, given the appeals that Glennan (2002) and Craver (2007) make to Woodward-style counterfactuals, one might suspect that the mechanical account will reduce to a manipulability account.

The mechanist can respond to this objection by appealing to the hierarchical character of mechanisms. Mechanisms consist of parts whose direct interactions are productive of some behaviour, but what counts as a direct interaction depends upon the level of analysis. In general, the parts of mechanisms are themselves mechanisms whose behaviours depend upon the organization and interaction of the parts of those parts. So for instance, one can explain the causal processes controlled by neuron circuits by invoking direct interactions between neurons at synapses. But the mechanism of synaptic transmission is itself a complex one, involving many constituent parts. These parts (e.g. vesicles, neurotransmitters, ions, and ion channels) are themselves mechanisms composed of parts whose interactions explain their behaviour. Thus, one can see that the change-relating generalizations—in this case a generalization describing how one action potential stimulates or inhibits another action potential—are explained by appeal to underlying mechanisms. These generalizations are *mechanically*

explicable (Glennan 1996: 61–3). Thus, the fact that such generalizations are appealed to in characterizing a causal relation does not undermine the mechanical character of the account.

The manipulability and mechanist accounts provide different criteria for identifying causal connections. On the one hand, there is the manipulability criterion—in this case, if one could manipulate the pre-synaptic neuron one would stimulate or inhibit the post-synaptic neuron. On the other hand, there is the criterion of an identifiable mechanism—in this case one has identified the mechanism of synaptic transmission. Is there a reason to think that one criterion more genuinely captures the nature of the causal relation?

Proponents of the manipulability theory will immediately point to what appears to be the Achilles heel of the mechanical account. Unless there is an infinite downward chain of nested mechanisms, sooner or later one will run out of mechanisms. One reaches a point where the parts of mechanisms interact in accordance with change relating generalizations that are not mechanically explicable. To take a simple case, imagine a Newtonian world in which two particles accelerate towards each other in virtue of their gravitational attraction. This attraction is characterized by a change-relating generalization describing the particles' changes in velocity, but there is not (we'll suppose) a mechanism that explains this. It is just a brute fact that massive objects *cause* other objects to accelerate toward them. Furthermore, the critic can make the following sort of reductive argument: because analysis of mechanically explicable generalizations at higher levels will ultimately bottom out in mechanisms whose parts interact in non-mechanically explicable ways, it will ultimately be the case that the causal character of such generalizations depends upon on what can be called fundamental or mechanically inexplicable generalizations. The virtue of the manipulability account is that it seems to work even at the fundamental level. While, at the fundamental level I cannot find a mechanism connecting X and Y, I can at least determine whether there is a causal connection by seeing if I can wiggle X to wiggle Y. In light of this objection, the manipulability theorist might argue that mechanists need manipulability but manipulability theorists can live without mechanisms.

But the advantage of the manipulability theory is less than one supposes. At some level the criticism of the mechanistic theory is that it cannot reductively eliminate causal concepts. But, as Woodward (2003) takes pains to point out, neither can the manipulability theory. At the root of the manipulability analysis is the transparently causal notion of an intervention. At the level of fundamental causal connections, how does one know that one has wiggled X (and just X)? One can have some confidence that one is wiggling X if one has an account of the wiggling mechanism, as well as an account of what mechanisms are used to isolate X from other factors that might be causally connected to Y. In the absence of such an account, the manipulability theorist has only the observation that the wiggling of X was followed by the wiggling of Y and the *supposition* that the wiggle of Y depended on that of X.

Minimally, the manipulability theorist has made an important point about the epistemology of causation. It is frequently the case that one can establish with reasonable certainty that a variable is causally relevant to another without knowing anything about the mechanism by which the variables are connected. Experimental manipulations can provide evidence *that* variables are connected, even in the absence of mechanical knowledge of *how* they are connected. But this epistemologically important point does not legitimate the manipulability theory as a metaphysical account of causation. To see this, consider the truth conditions for

the interventionist counterfactuals that underlie the manipulability theory. Do these truth-conditions depend on singular or general facts? Woodward's answer is clear. Although causal claims are often made at the type level, the truth of these claims depends upon the truth of claims about individuals: '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 individuals will change the value of Y located in particular individuals' (Woodward 2003: 40). So, to understand the truth-conditions for type-level claims, we must understand the truth-conditions for the token claims. These truth conditions are specified in terms of interventionist counterfactuals. If one were to intervene on (this instance of X to change its value) one would change Y. But how are we to understand the truth-conditions for these counterfactuals? Suppose at some time t I flip a switch (e_1) and a light goes on (e_2). As an epistemic matter, I establish that e_1 caused e_2 by manipulating the switch at other times t', but the truth of the causal claim at t does not depend upon the truth of correlations at t'. It depends upon the singular counterfactual dependency at t.

Stathis Psillos (2004) has raised this concern about Woodward's account, arguing that Woodward has given evidence conditions for interventionist counterfactuals, but that he hasn't given truth-conditions. Psillos suggests that the appropriate way to correct this deficiency is to appeal to laws of nature. Although he does not spell out how this appeal would work in detail, roughly it must be that the counterfactual dependency would be true in virtue of the fact that the sequence of antecedent and consequent was a tokening of a lawful generalization. To make this proposal work, one needs an account of laws of nature, and Psillos endorses the Mill–Ramsey–Lewis (MRL) approach, in which laws are taken to be generalizations that are theorems (or axioms) of a deductive system that provides the best combination of simplicity and strength (Lewis 1973: 73).

In adopting this view, Psillos has just traded one problem for another. The MRL account of laws is a Humean position that grounds the truth of laws in the totality of particular facts in the actual world (cf. Ch. 7). If truth-conditions for counterfactuals are grounded in MRL laws, Psillos's proposal does provide truth-conditions for counterfactuals that are grounded in epistemically accessible facts about the actual world, but then the facts upon which a singular counterfactual (like the fact that if I were to flip the switch I would turn on the light) depends are facts not just about the particular instance but about the whole class of particulars. Thus, it violates the principle that the truth-conditions for singular causal claims should be intrinsic, that whether a change of an instance of X causes a change of an instance of Y should depend only upon the local facts surrounding these instances. Thus Woodward should rightly reject Psillos's attempt to ground the truth of singularist interventionist counterfactuals in MRL laws.

Except in the case of fundamental laws, the mechanical theory of causation can provide the truth-conditions for interventionist counterfactuals that the manipulability theory seems to lack. When causal relations are mechanically explicable, what makes it the case that wiggling some X will produce the wiggling of some Y is that there is an intervening mechanism between X and Y.[3] Understanding the nature, structure, and functional organization of the parts that make up that mechanism will allow one to determine the range of counterfactual circumstances under which the dependency between X and Y would be maintained—roughly those circumstances in which the mechanism will not break down.

Given that any variables connected by intervening mechanisms depend on mechanically

inexplicable causal connections, it does seem incumbent on the mechanical theorist to say something about how one can know that such connections obtain. The mechanist supposes that causal connections are constituted by interactions between parts of mechanisms, and that the interactions between parts of mechanisms may, on further analysis, be constituted by nested mechanisms, but that ultimately one will bottom out with parts that interact, where these interactions aren't mechanically explicable. Suppose that one observes that changes in one part are followed by changes in another part—perhaps changes that occur without intervention or perhaps changes that one believes to be caused by an intervention. How does one know that the changes in one part really *produce* the changes in the other part? This is Hume's problem reproduced at the level of the most basic constituents of mechanisms. Following the MRL approach, one might take the fundamental laws to be those change-relating generalizations that give us the simplest and strongest account of the facts in the world. Neither the mechanist nor the manipulability theorist needs to suggest that fitting within the strongest and simplest account is what makes it true that these variables are causally connected, but the mechanist can say that this is the best reason we have, at the fundamental level, to accept certain generalizations as expressing causal relations.

3. ACTIVITIES AND INTERACTIONS

MDC's analysis of mechanisms is distinguished by its introduction of the concept of an activity. MDC characterize mechanisms as 'entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions' (MDC 2000: 3). MDC claim that the introduction of the term 'activity' reflects a crucial metaphysical innovation. They argue for this claim by characterizing the debate over the nature of mechanisms in terms of a dispute between three ontological stances. First there are the substantivalists (including Bechtel and Richardson 1993 and Glennan 1996; 2002) whose basic ontology is an ontology of entities. 'Substantivalists confine their attention to entities and properties, believing it is possible to reduce talk of activities to talk of properties and their transitions' (MDC 2000: 4). Second, there are process ontologists (notably Rescher 1996), who 'reify activities and attempt to reduce entities to processes' (ibid.). They describe their new third way as 'dualist', requiring both entities and activities as ontologically irreducible categories.

MDC appear to overstate the novelty of their position because they fail to recognize the dualist character of the substantivalist account. Where MDC speak of entities and activities, Bechtel, Richardson, and Glennan speak of parts and interactions. No account of a mechanism can get along talking about the entities alone, but for Glennan and for Bechtel and Richardson this additional ingredient is provided by interactions.

MDC argue that this language is insufficient: '[I]t is artificial and impoverished to describe mechanisms solely in terms of entities, properties, interactions, inputs–outputs, and state changes over time. Mechanisms do things. They are active and so ought to be described in terms of the activities of their entities, not merely in terms of changes in their properties' (ibid. 5). MDC equate interactions with state changes over time. What this suggests is that the real target of their criticism is a Humean conception of interactions in which interactions are

characterized 'merely in terms of changes in ... properties'. Their view is that productivity cannot be reduced to mere change in properties. Given the manifold problems with Humean approaches to causation, this is certainly a respectable position to take, but neither Glennan nor Bechtel and Richardson adopt the Humean approach. If activities can be productive, so can interactions.

Machamer (2004) and Bogen (2005) have recently offered more detailed defences of the metaphysical significance of activities. The crux of their defence is that activities provide a better and more natural approach to understanding productivity than do interactions, because activities can help us understand productivity without appeal to counterfactuals.

First consider how counterfactuals appear to give some understanding of the productive character of interactions. When a change in one part of a mechanism produces a change in another part of a mechanism, the claim that the first change *produces* rather than just precedes the second change is (ignoring a lot of details) taken to be licensed by whatever evidence one has for the counterfactual claim that if the first change had not occurred, neither would the second. The modal character of the productivity is cashed out in terms of the counterfactual.

Bogen worries, however, that this analysis makes the truth of a causal claim depend upon what would have happened in other circumstances—what he calls counterfactual generality—rather than upon what actually does happen. At root, both Bogen and Machamer are appealing to a well-known argument from Elizabeth Anscombe:

If A comes from B, this does not imply that every A-like thing comes from some B-like thing or set-up or that every B-like thing or set-up has an A-like thing coming from it; or that given B, A had to come from it, or that given A, there had to be B for it to come from. Any of these may be true, but if any is, that will be an additional fact, not comprised in A's coming from B. (Anscombe 1993: 92)

Anscombe insists that the production of a particular B from a particular A is independent of any actual or counterfactual regularity. Bogen and Machamer seem to think that activities capture the notion of productivity without reference to any such regularities.

I will not dispute the claim that the productive continuity between cause and effect is a fact about the actual world in the particular case. What I will dispute is that Anscombe, Machamer, or Bogen have adequately explained what this productivity is. Anscombe writes 'causality consists of the derivativeness of an effect from its causes. This is the core, the common feature, of causality in its various kinds. Effects derive from, arise out of, come of, their causes' (ibid.). MDC (2000: 4) echo this when they characterize activities: 'Activities are the producers of change. They are constitutive of the transformations that yield new states of affairs or new products.' But these characterizations of activities and causes do little more than offer a set of synonyms for 'cause'. They hardly elucidate the concept of cause or production. Neither Anscombe nor MDC would worry too much about this, because they believe that we acquire our understanding of causation by starting with the understanding of particular activities—pushing, bending, hitting, etc. But while this is certainly true, to the extent that our concept of cause does mean something, we must give some account of what it means in general, and to provide a synonym for 'cause', like 'produce', is not to provide an

analysis. The virtue of the manipulability approach to causation is that it tells one something quite general about causes—that they can be used, in an idealized way—to manipulate effects. The virtue of the mechanical approach is again that it tells one something quite general about causes—that causes and effects will generally be connected by intervening mechanisms. Neither of these theories provides reductive analyses of causation, but both say something non-trivial about the nature of causation.

4. MECHANICAL SYSTEMS AND MECHANICAL PROCESSES

While the term 'mechanism' is currently most commonly associated with the concept of mechanism identified with the work of Bechtel, Glennan, and Machamer, Darden, and Craver, the term 'mechanism' first entered into the contemporary discussion of causation and explanation with the work of Salmon and Railton, and the 'causal-mechanical' approach to explanation (Railton 1978; Salmon 1984). What distinguishes the earlier Salmon/Railton approach from the more recent approach is that Salmon and Railton conceive of mechanisms as a network of interacting *processes*, whereas the more recent mechanists think of mechanisms as *systems*—organized collections of parts.

To help understand the difference between the approaches, consider paradigm cases of each sort of mechanism. For a systems theorist a paradigm case might be a toilet, while for the process theorist a paradigm case might be a baseball striking a window. The toilet is a *thing*—a structured system consisting of parts (valves, levers, floats, etc.) that interact in regular ways. A baseball striking a window involves a process which involves a series of events (the pitch of the baseball, the hit, the collision with the window, etc.), but we can't think of this sequence of events as a thing. The *operation* of system mechanisms gives rise to processes (e.g. toilets flush), but these processes are, in virtue of the stability of the mechanism, regular and repeatable.

By attending to the distinction between these two kinds of mechanisms, we can see a significant limitation in the mechanical approach to causation proposed in Glennan 1996. The hope expressed in that paper was that any two causally connected events that were not connected in virtue of a fundamental (i.e. mechanically inexplicable) law would be connected by the operation of an intervening mechanism. If mechanisms are construed as systems as they are in the accounts of Glennan, Bechtel, and MDC, it is evident that there are many true singular causal claims in which a cause is not connected to its effect via the operation of a mechanism qua system. When the baseball breaks the window, there is a causal process that leads from the pitch of the baseball to the breaking of the window, but there is no mechanism qua system for window-breaking.

Clearly one of the attractions of the process approach is that it appears to give an account of causal connectedness for singular causal sequences. But, it does so at a cost. As Hitchcock (1995) has argued, process theories appear unable to provide any intuitive understanding of the explanatory relevance of causes for effects. Relatedly, a process theory that identifies interactions in terms of exchange of conserved quantities does not provide informative and explanatory accounts of the nature of interactions between parts at higher levels of organization (Glennan 2002; Psillos 2004).

What appears to be required to provide a more adequate and general mechanistic account of singular causal sequences is to describe causal processes in terms of the interactions of parts, as is done in the systems approach, but to recognize that in many singular causal chains, the parts are not organized in a stable system. Thus for instance, one can conceive of the baseball sequence in terms of the interaction of a set of parts (baseball, bat, and window), but these parts are not arranged in any stable configuration. Consequently, the next pitch won't necessarily be hit, and if hit, it probably won't travel on exactly the same trajectory. But what one does know is that when bats hit balls in certain ways (i.e. with certain velocities and spins) they alter trajectories in reliable ways that can be described by 'change-relating generalizations'. Thus we see the striking of a ball in a certain way as explanatorily relevant for the changing of that ball's velocity and ultimately for the breaking of the window. The singularity of the causal sequence arises from the fact that the particular arrangement and state of the parts of the mechanism on a given occasion are ephemeral, but the parts themselves and their behavioural dispositions are robust. An account of such 'ephemeral mechanisms' is briefly suggested in Glennan 2002, but much work would be needed to show how such an approach could improve upon existing accounts of singular causation.

FURTHER READING

For a general introduction to recent work on mechanism it is good to begin with Glennan (1996) and MDC (2000). The former work is more focused on issues of causation, while the latter discusses a range of topics, including models, explanation, and discovery. Bechtel and Richardson (1993) is a book-length treatment of approaches to discovering mechanisms, especially in biological systems, but it anticipates much of the general mechanistic approach developed over the last decade. It is helpful to contrast this view of mechanism with the theory of causal processes discussed in Ch. 10 above.

Concerning activities, the reader might begin with Machamer's (2004) defence. Tabery (2004) argues that both MDC's activities and Glennan's interactions are necessary for a proper account of mechanisms. Psillos 2004 provides a useful analysis and critique of activities as well as of Glennan's approach, and seeks to provide an account of the relationship between mechanistic and manipulationist and counterfactual accounts of causation.

There has been considerable work done in recent years on developing a theory of causal mechanical explanation. In the 1980s and 1990s, the term 'causal mechanical explanation' was chiefly applied to Salmon's (1984) approach. Glennan (2002) contrasts Salmon's approach to one consistent with the systems approach to mechanism. Thagard (1999) provides a related account of mechanisms and mechanistic explanation in the context of medical science. Bechtel and Abrahamsen (2005) offer a general account of mechanistic explanation, contrasting it with nomological approaches. Using case studies from neuroscience, Craver (2007) provides a comprehensive theory of mechanistic explanation. Craver's book also has substantial discussions of the relationship between mechanisms, manipulability, and causation and of the implications of the mechanistic approach for reduction and problems of higher-level causation.

REFERENCES

ANSCOMBE, G. E. M. (1993). 'Causality and Determination', in Ernest Sosa and Michael Tooley (eds.), *Causation*. Oxford: Oxford University Press, 88–104.

BECHTEL, WILLIAM, and ABRAHAMSEN, ADELE (2005). 'Explanation: A Mechanist Alternative', *Studies in the History and Philosophy of Biology and Biomedical Sciences* 36/2: 421–41.

—— and RICHARDSON, ROBERT C. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

BOGEN, JIM (2005). 'Regularities and Causality: Generalizations and Causal Explanations', *Studies in the History and Philosophy of Biology and Biomedical Sciences* 36/2: 397–420.

CARTWRIGHT, NANCY (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

CRAVER, CARL F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. New York: Oxford University Press.

DIJKSTERHUIS, E. J. (1969). *The Mechanization of the World Picture*. London: Oxford University Press.

DOWE, PHIL (2000). *Physical Causation*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge: Cambridge University Press.

GLENNAN, STUART S. (1996). 'Mechanisms and the Nature of Causation', *Erkenntnis* 44/1: 49–71.

—— (2002). 'Rethinking Mechanistic Explanation', *Philosophy of Science* 69/3 suppl.: S342–S353.

HITCHCOCK, CHRISTOPHER R. (1995). 'Discussion: Salmon on Explanatory Relevance', *Philosophy of Science* 62/2: 304–20.

LEWIS, DAVID K. (1973). *Counterfactuals*. Oxford: Blackwell.

MACHAMER, PETER (2004). 'Activities and Causation: The Metaphysics and Epistemology of Mechanisms', *International Studies in the Philosophy of Science* 18/1: 27–39.

—— DARDEN, LINDLEY, and CRAVER, CARL F. (2000). 'Thinking about Mechanisms', *Philosophy of Science* 67/1: 1–25.

MITCHELL, SANDRA D. (1997). 'Pragmatic Laws', *Philosophy of Science* 64/4: S468–S479.

PSILLOS, STATHIS (2004). 'A Glimpse of the Secret Connexion: Harmonizing Mechanisms with Counterfactuals', *Perspectives on Science* 12/3: 288–319.

RAILTON, PETER (1978). 'A Deductive-Nomological Model of Probabilistic Explanation', *Philosophy of Science* 45: 206–26.

RESCHER, NICHOLAS (1996). *Process Metaphysics: An Introduction to Process Philosophy*. SUNY Series in Philosophy. Albany: State University of New York Press.

SALMON, WESLEY C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

TABERY, JAMES G. (2004). 'Synthesizing Activities and Interactions in the Concept of a Mechanism', Philosophy of Science 71/1: 1–15.

THAGARD, PAUL (1999). *How Scientists Explain Disease*. Princeton, NJ: Princeton University Press.

WOODWARD, JAMES (2000). 'Explanation and Invariance in the Special Sciences', *British*

Journal for the Philosophy of Science 51/2: 197–254.

—— (2002). 'What is a Mechanism? A Counterfactual Account', *Philosophy of Science* 69/3 suppl.: S366–S377.

—— (2003). *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

CHAPTER 16
CAUSAL PLURALISM

PETER GODFREY-SMITH

1. INTRODUCTION

Causal pluralism is the view that causation is not a single kind of relation or connection between things in the world. Instead, the apparently simple and univocal term 'cause' is seen as masking an underlying diversity.

Assessing such a claim requires making sense of a difficult counting operation. How do we tell whether a theory of causation is identifying causation with a 'single' kind of connection? In practice, there tends not to be much disagreement about how to do the counting, because most philosophical work on causation has sought a view with an obvious kind of unity. The literature often works with a standard range of *candidate* connections that seem to have an important link to the idea of causation. These include natural laws, de facto regularities, counterfactual dependence, probabilistic dependence, and some others. It has been common in philosophy to pick one of these and try to make sense of causation entirely in those terms. The candidate chosen is seen as either fundamental to our thinking about the world, a fundamental ingredient of the world itself, or both. Against that background, it is an unorthodox move to say that all such unified treatments of causation are mistaken; causation is, with respect to the ingredients recognized as distinct options in philosophical practice, irreducibly plural or diverse.

I am grateful to Alison Gopnik, Thomas Hofweber, and Francis Longworth for helpful discussions of these issues.

Appropriately, causal pluralism comes in several kinds. The main focus of this survey is a collection of recent treatments that directly oppose the kind of unity seen in traditional analyses. The simplest kind of pluralism treats our ordinary talk of causation as shifting between two distinct 'concepts' of cause (sect. 3). This view posits an ambiguity that could be resolved if each concept was given a different name. Another suggestion is that we have a single concept of causation, but one whose use is guided by what Brian Skyrms called an 'amiable jumble' of criteria that sometimes work together and sometimes pull apart (sect. 4).

Those two options use traditional philosophical raw materials to make sense of causation (regularities, counterfactuals, etc.) but put them to work in unorthodox ways. A more unusual style of analysis is seen in a proposal by Anscombe. Perhaps 'cause' is a generic term whose application to the world is parasitic on the semantic properties of a range of terms (such as 'burn' and 'scrape') that pick out specific kinds of physical connection. Anscombe's proposal

raises the possibility of a kind of 'minimalism' about causation, and also a 'family resemblance' view that is closely related (sect. 5).

After surveying these options, I suggest that an understanding of the workings of our concept of causation may require that we treat it as something like what W. B. Gallie (1956) called an 'essentially contested' concept. These are concepts whose role in our conceptual scheme is pivotal in a way that makes them permanently resistant to definition and the drawing of stable boundaries.

2. Methodological and Other Preliminaries

When someone says causation is really 'two kinds of thing', or a vague jumble of things, are they making a claim about our *concept* of causation, or a claim about the real-world phenomena that causal talk is *directed* on? They might be making either claim, or both. The literature does not always distinguish between the options here.

Usually, the main focus is on 'our concept' of causation. A pluralist might claim that, despite surface appearances, our talk of causation is guided on different occasions by different sets of criteria. In most cases, this will have the consequence that the worldly connections that count as causal will themselves be disunified (because they satisfy different criteria). But in principle, pluralism at the level of concepts allows the possibility that, at least in the actual world, all the cases that satisfy one set of criteria also satisfy the other. This correlation might even hold as a matter of natural law. In such a case, causation would be unified in the world but not in our thinking. (Compare Hesperus and Phosphorus, heat and molecular motion.)

Is the converse also possible? That is, perhaps we apply a univocal test when looking for causation, but the phenomena it picks out are, in some deep sense, plural in character. If so, causation would be unified in our thinking but not in the world. (Compare the case of *jade*, at least according to the philosophers' story in which jade is two different minerals that we lump together into one category without realizing it.) Views of causation that have this structure might seen natural to Cartesian dualists, who may say that physical and mental causation are deeply different phenomena that we get an informal handle on via a single abstract set of criteria. When this is labelled as a kind of 'causal pluralism', it might be objected that if all the cases do satisfy one set of abstract criteria, then in one important sense causation is unified in the world as well.

In this survey I will try to be explicit about whether claims are being made about our words and concepts, on one side, or the phenomena that our thought and talk are aimed at, on the other. The total picture is something like this. We can distinguish three relevant sets of facts. One is the total set of facts about our habits of thought and talk in this area—how ordinary people, scientists, doctor, lawyers, and so on, use the word 'cause' and its relatives, how responsibility is attributed and how explanations are given. A second set of facts concerns what the world contains and how the world runs. A third set of facts concerns the relations *between* the other two. For example, does the ordinary concept of cause succeed in picking out a real 'natural kind'? Causal pluralism in its familiar forms usually makes claims about the first set of facts, those about causal thought and talk. But the other sets of facts often become relevant, and it is important to keep an eye on exactly what sort of thing an alleged plurality is

supposed to be.

A couple of other distinctions between pluralisms need to be made before moving on. The first is subtle. There are usually two steps in mainstream analyses of causation. Speaking metaphorically, one step is finding the right raw material, and the other is building the right structure out of it. We might decide that causation is essentially a matter of counterfactual dependence, but we then have to work out which kinds of counterfactual dependence suffice to make C a cause of E. Even bracketing the possibility of pluralism about the raw materials, philosophers may have often assumed too unified a view about how the raw materials should be put together (Hitchcock 2003). The usual target for philosophical analysis is what it is for one thing to *be a cause of* another. But there might be a family of causal concepts, including triggering, enabling, hastening, delaying, being linked in a causal chain ... Once we have worked out whether C hastened E, perhaps it is pointless to work out whether this is also enough for C to *be a cause of E*. Hastening is just what it is, and it is one genuine causal relation. The focus below will be primarily be on the question of pluralism at the level of raw materials, but sometimes the two questions may interact.

Lastly, there are some ways in which one might assert a plurality of causes that are not important here. Most views of causation will hold that any event will have multiple causes (e.g. more immediate and more remote causes). This is compatible with each cause being a cause in the same sense (e.g. a counterfactual sense), so it does not amount to pluralism of the kind under discussion. Secondly, an analysis of causation might make use of a *conjunction* of very different criteria. This usually counts as a univocal treatment of causation, unlike a disjunctive treatment, even though highly heterogeneous conjunctions should probably seem philosophically puzzling as well.

3. TWO CONCEPTS OF CAUSE?

The simplest form of causal pluralism, and the one that has been developed in most detail, is the idea that there are two distinct *concepts* of cause that have become tangled together in our language. We have one term, 'cause', that tends to be guided by two different sets of criteria that pull apart in some cases. This kind of causal pluralism could be fixed, in principle, by a relabelling. If we began speaking of 'cause$_1$' and 'cause$_2$', no trouble would remain. Each of these concepts could be assessed in its own right, with respect to how it relates to the world. Each might pick out a natural kind, one might be empty, and so on. This low-key pluralism treats the word 'cause' as analogous to ordinary examples of ambiguity, such as 'bank' or 'bat'. 'Two concepts' proposals have been outlined in detail several times (Sober 1985; Hall 2004), and there have been a lot of informal suggestions along the same lines (always apparently with $n = 2$).

The best worked-out view of this kind is seen in Hall (2004). Hall focuses on a particular family of tensions in many people's causal intuitions, which are often, and reasonably, taken to motivate a pluralist view of some sort or other (see also Hitchcock 1998; Schaffer 2000). I will refer to these as intuitions based on *difference-making* and *production* as rival marks of causation. Sometimes a factor with the right role in the physical production of an event will be classified as causal, whether or not it made a difference to the outcome. But sometimes what

matters is whether the factor made a difference. These criteria can pull apart even in everyday cases. In cases of *redundant* causation we see the role of production intuitions. The bullet strikes its victim and produces his death. This is true regardless of whether other or not other bullets were right behind it, ready to produce the same effect. Being suitably located in a producing mechanism suffices for causation, apparently regardless of difference-making.

In cases of *causation by prevention* we see the role of difference-making intuitions. The air traffic controller sees the two planes on collision course, but you prevent him from issuing a warning. Your action would usually be considered one cause of the ensuing crash. This is true despite the absence of a physical connection between your action and the effect. You made a difference to how things went, and difference-making suffices for causation in these cases.

Hall argues that the tension between production-based criteria and difference-making extends beyond intuitions about particular cases, to affect general principles. For example, many philosophers have wondered whether causation is transitive. Focusing on production makes it seem that the answer is yes; focusing on difference-making suggests that the answer is no. Hall's response to the situation is to say that we have two distinct concepts guided by different sets of criteria. There is causation in the sense of production, and causation in the sense of difference-making. (Hall actually calls this one 'dependence'.) If we distinguish them, we find that many of the plausible general principles about causation apply to one kind of relation or the other, but not to both. For Hall and others, difference-making is naturally analysed using counterfactuals, in the tradition of Lewis (1973), while production is harder to characterize.

In more recent work (unpublished), Hall has expressed doubts about the 'two concepts' analysis. This is because there are cases where a factor *C* looks like a cause of *E*, despite passing neither of his two tests. These are cases where *C* is a *redundant threat-canceller*. *C* blocks a threat *T* that would otherwise have prevented *E,* but if *C* had not acted then *D* would have sufficed to block *T* instead. A threat-canceller such as *C* can lack any physical connection to *E* (think of the air traffic control case), and in this case *C* was not a difference-maker either. So *C* passes neither test but might reasonably be said to be a cause of *E*. (Here the reader may, with good reason, consider applying the ideas from Hitchcock sketched at the end of the previous section. Maybe there is no need to ask the standard ' ... was a cause of ... ?' question in such a case.)

So the 'two concepts' view has internal problems. And once we have moved into the pluralist camp, it is worth asking whether this approach is the most promising in general. If it were really the case that people had two concepts, then actual usage should contain certain kinds of evidence of this. We should not merely see attempts to *clarify* causal claims, but attempts to *switch* the hearer from one sense of 'cause' to another. A good model would be provided by words such as 'mad' and 'funny' which have two senses, related to each other but distinct, and which work in ways that sometimes make disambiguation requests appropriate. ('Did you mean it was funny as in *weird,* or as in *ha-ha*?') But this switching does not seem visible in causal discourse; there the situation seems more disorderly. That motivates a second kind of pluralism.

4. THE AMIABLE JUMBLE

In 1984 Brian Skyrms, in a short paper about quantum mechanical puzzles, suggested that ordinary causal description is guided by an 'amiable jumble' of criteria. These criteria usually work together in macroscopic cases, but can pull apart in special situations, especially those unearthed by physics. Causal thinking is a rough-and-ready framework, well suited to the everyday world, but not something that helps us describe the fundamental processes that make the world run.

The criteria Skyrms mentions as part of our 'jumble' are familiar tools for analysing causation, including spatio-temporal connection and instantiation of a regularity. Such a view will naturally add that there is no fact of the matter about which of these tests are the central ones; the evolution of the concept has not been of a kind that has made such a prioritization necessary. Different people are free to weight different tests differently, and free to use different weightings on different occasions. So the idea might be summarized thus:

> AMIABLE JUMBLE: '*C* was a cause of *E*' is true iff the relation between *C* and *E* satisfied some contextually appropriate combination of our amiably jumbled criteria for causation (instantiating a regularity, being spatio-temporally connected in certain ways, inducing counterfactual intuitions ...).

Skyrms himself thought that the criteria tend to work together in everyday contexts, as noted above. His article was written prior to the focused work on counterfactual dependence that has revealed ways in which different criteria can pull apart in everyday cases (such as cases of causation by prevention). It now seems that the jumble is not nearly as amiable as Skyrms envisaged; perhaps 'cantankerous jumble' would be a more appropriate term.

This is a simple but sensible proposal, and related views have been defended by other writers. It is sometimes unclear which tools from the philosophy of language are most appropriate here, and how they are related. For example, how does the amiable jumble view relate to the idea that causation is a 'vague' concept? Is this the same as claiming that causation is a 'cluster concept'? What about the idea of 'family resemblance'?

Many cases of vagueness have nothing to do with pluralism in the sense under discussion, because they involve the vague application of a single test, often a one-dimensional gradient standard. (How many hairs can one have while still counting as bald?) But it seems reasonable to say that *one* way to get a kind of vagueness is for a concept to be guided by a plurality of jumbled criteria that can pull in different directions. In the case of causation, many components of the cluster may themselves be vague.

Secondly, I take the amiable jumble proposal to be essentially the same as the claim that causation is a 'cluster concept'. Richard Healey, responding to the same quantum mechanical puzzles as Skyrms, proposed in 1994 that causation is a cluster concept, and the underlying picture is the same. These writers do not say much about how exactly a jumbled or cluster concept works. Longworth (forth-coming) remedies this with a detailed treatment of cluster concepts and their relation to causation. Modifying his account, we might say that a cluster concept is one whose application is guided by a set of distinct criteria where (1) none of the criteria are individually necessary, (2) the entire set is clearly sufficient, (3) at least one of the

criteria must be met, and (4) some proper subsets are sufficient, though many of these subsets generate marginal or contested applications.

The cluster-concept or amiable jumble view seems to do better than the two-concepts approach in accommodating everyday facts about the use of causal language. (Here I rely on informal observation, not on empirical data.) In particular, the cluster concept view does not predict that it should be possible to switch a hearer from one discrete 'sense' of causation to another. What we might expect to see instead is a practice of guiding interlocutors to give more weight to some criteria and less weight to others, and an occasional willingness to give stipulative precisifications ('Well, what I mean by "cause" here is ...'). This is, I suggest, what we do tend to see.

If causation is a cluster concept, is this a case of 'family resemblance'? In many ways this term seems fair, and Longworth (forthcoming) associates the two ideas closely. However, I reserve the term 'family resemblance' for a more unusual option.

All the views discussed so far have drawn on *abstract* criteria that can be associated with causation. These include 'instantiating a regularity', 'being linked in a chain of counterfactual dependence', and so on. These criteria are not based on the particular physical character of a connection between two things. This might seem natural, as it seems unlikely that all the physical (and perhaps non-physical) connections that are recognized as causal by ordinary criteria will have similarities in their intrinsic make-up and structure. That fact usually pushes the philosopher towards a more abstract treatment. But this push is not so forceful once a pluralist attitude is on the table. This brings us to the possibility of a more unorthodox form of analysis.

5. ANSCOMBE, MINIMALISM, AND FAMILY RESEMBLANCE

Elizabeth Anscombe's 'Inaugural Lecture' on causation (1971) raised, in a few striking paragraphs, the possibility of an analysis in which a range of familiar terms that we use to pick out various specific causal relations—such terms as 'scrape' and 'burn'—are *semantically prior* to the general term 'cause'.

The word 'cause' itself is highly general. How does someone show that he has the concept *cause*? We may wish to say: only by having such a word in his vocabulary. If so, then the manifest possession of the concept presupposes the mastery of much else in language. I mean: the word 'cause' can be *added* to a language in which are already represented many causal concepts. A small selection: *scrape, push, wet, carry, eat, burn, knock over, keep off, squash, make* (e.g. noises, paper boats), *hurt*. But if we care to imagine languages in which no special causal concepts are represented, then no description of the use of a word in such languages will be able to present it as meaning *cause*. (ibid. 93)

The last claim—that no word could mean *cause* in a language with no 'special causal concepts'—is the key one. Anscombe, as I interpret her, is saying that the essential semantic role of the word 'cause' in languages such as English is to collect together a range of more specific relations that we also pick out with such words such as 'squash', 'push', and so on.

The application that the word 'cause' has to phenomena in the world goes *via* the application that these more specific terms have. If a language lacked any words such as 'scrape' and 'push' there would be nothing in the language for the general term 'cause' to collect together.

Anscombe did not give much more of an analysis than this, but one way to make the proposal more precise is to develop it as a kind of 'minimalism' about causation.

MINIMALISM:

1. '*C* was a cause of *E*' is true iff the relation between *C* and E can also be described using some member of set *S*, or can be described as a chain of relations each of which can be described using some member of *S*.

2. *S* is a set of causal verbs and other linguistic formulas which represent 'special causal concepts' in Anscombe's sense.

Initially, for the purpose of exploring the simplest version of such a view, we can imagine that the composition of set *S* is taken as unexplained. All we know is that some specific relations are grouped together by the word 'cause'.

The term 'minimalism' is intended to suggest an analogy with minimalism about truth, which holds that the meaning of 'true' is captured entirely by the T-schema: '"*p*" is true iff *p*' (Horwich 1990). For the minimalist, the word 'true' is a device for adding and removing quotation marks, thereby making possible the compact expression of various claims that would otherwise require a convoluted or infinitely long sentence. The word 'true' does not pick out some special matching or correspondence relation between signs and the world. Similarly, causal minimalism holds that to assert that some connection is causal is not to attribute some special empirical or modal feature to it. To call a connection causal is to say that this connection is one that can also be described using some unspecified member of *S*. We use the term 'cause' to generalize and abstract, to form sentences such as 'Bill caused a lot of trouble', which have the same sort of role that 'Everything Bill believes is true' has on a minimal theory of truth.

I do not know of an explicit defence of a minimalist view of this kind. Cartwright (1999; 2004), who draws on Anscombe, may be close to it, but her view may instead be closer to one of the other options discussed below.

Minimalism is philosophically intriguing, but has clear problems. First, it is significant that Anscombe's examples of 'special causal concepts' describe relations between objects, not relations between events or facts. So the operation of abstraction that is achieved by using the word 'cause' is not as simple as what we see in the case of minimalist theories of truth. It is hard to tell how damaging this is to the basic thrust of minimalism. A more important problem concerns the possibility of discovering novel kinds of causal relation. According to minimalism, no relation could be truly called causal if we had not previously developed a more specific description for that type of connection and grouped it as causal. So, for example, if minimalism was true it would make no sense for a person in seventeenth-century England to wonder whether the Great Fire of London had somehow caused the end of a plague epidemic

that had preceded it, while having no idea of any mechanism by which this could have happened.

Can this problem be fixed while remaining within the spirit of Anscombe's view? It is at this point that the notion of family resemblance may be useful. Perhaps the relations grouped as causal are unified by a set of similarities between the relations themselves—between burning, pushing, scraping, and so on. These similarities might be real even if there is no abstract test that can be used to describe all these relations (in terms of regularities or counterfactuals, for example). So when someone wonders whether some coincidence or sequence has an unknown causal relation underlying it, they are wondering whether there is *some* relation between the two events that, once understood, will be found to have a family resemblance to the familiar cases of causal relations.

Here I treat the family resemblance view as different from the amiable jumble or cluster concept view. The cluster concept view supposes that we apply a jumble of abstract tests; the family resemblance view appeals to similarities between the characteristics of the specific relations themselves.

To the extent that this is a distinct option, does it handle the phenomena of causal thought and talk as well as the amiable jumble view does? If we look at how people think about complicated causal relations, then the cluster view seems more promising. It seems clear that in many cases people do engage in an abstract assessment of difference-making, for example. This is so abstract a mode of assessment that it is hard to make sense of within an Anscombe-inspired approach. Recent empirical work suggests that abstract assessment of difference-making might be quite psychologically deep in us as well (Gopnik and Schulz 2004). On the other hand, there are simple observation-guided forms of everyday causal description which seem to be handled very naturally by Anscombe's view or something like it. I can *see* you squash the pie, and that is enough for me to say that you caused various things. This suggests that the right analysis might include a role for a cluster of abstract tests, *and* a role for family resemblance at the level of the specific causal relations themselves.

To some this last move will also seem like a 'kitchen sink' sort of option, philosophically. Is *anything* left out? Surely the pluralist urge can go too far. It can, but perhaps what we should take from the preceding discussion is the idea that a cataloguing of tensions and complexities in our concept needs now to be accompanied by a different kind of analysis.

6. CAUSATION AS AN 'ESSENTIALLY CONTESTED' CONCEPT?

The idea of an 'essentially contested concept' (ECC) was introduced by W. B. Gallie (1956). Gallie's aim was to describe how some concepts remain permanently resistant to definition and analysis because of the pivotal place they occupy in our conceptual scheme. These are cases where it is not just hard to work out when the conditions of application of a term are met, but cases where the conditions for application themselves are, given the concept's role, permanently susceptible to being challenged and renegotiated.

I suggest that something like this analysis might be applied to the case of causation. The details of Gallie's analysis, which was designed to deal with such cases as *art* and *democracy*, do not fit the case of causation very well. But his treatment provides a good starting point.

As Gallie originally conceived them, ECCs are used to appraise human activities and achievements. There are accepted paradigm cases that exert influence on how the term is used, but the complexity and changeability of the domain is such that there will be no obvious and undeniable rules for extending the application of the term to new cases. An ECC is used with evaluative loading, and there will typically be consequences when it is successfully applied. As a result, the term's normal use occurs against a background of recognized dispute over its application. Proper use of the term itself involves a contesting of other uses, and recognition that one's use will in turn be contested.

The concept of cause is not in this category. It is (mostly) used to describe relations between events or facts, not to appraise human activities. It also seems too strong to say that a recognition of the ongoing contesting of criteria is a background assumption of normal use. But the concept of cause has become embedded in our practices in such a way that this wrangling is very *likely*. And in some kinds of causal discussion, the concept can have the 'aggressive' character that Gallie was interested in. So 'cause' (unlike 'art' and 'democracy') has low-level uses that are intended to be uncontentious, but it also has a richer ECC-like role in some contexts.

We might then treat causation as a more low-key relative of an ECC. It is a concept that will be reliably subject to dispute with respect to its boundaries and criteria for application. We expect terms to acquire this role when their successful application has significant downstream consequences, but their domain is complex in ways that involve the absence of sharp borderlines that function as attractors to usage. As in the case of Gallie's ECCs, an accepted set of exemplars and a sense of a shared *purpose* behind diverse uses prevent a fragmentation into distinct concepts.

These ideas might be linked to tools developed in recent 'inferentialist' philosophy of language (Brandom 2000). Terms such as 'freedom' and 'cause' tend to have significant consequences when successfully applied, but complicated criteria for application. The result is frequent challenge to whatever 'language entry rules' might be operating.

The concept of causation has these features as a consequence of its role in a family of important practices. Perhaps the crucial one is the assignment of responsibility. When it is established that a person is causally responsible for some event, they are often subject to praise, blame, and sanction. Within the standard set of problem cases in philosophy, cases of causation by omission are particularly relevant (Beebee 2004). Suppose you walk past a child who has fallen in a pond, who then drowns. We *might* say that you can be held responsible for the child's death even though (because of the absence of physical connection) you were not a cause of it. That is, we might insist that there can be a non-causal basis for moral responsibility. But there is a great deal of pressure in the other direction. It becomes more unproblematic to hold you responsible if we treat your act as a cause.

So Gallie's idea of an ECC, or a modification of it, might have useful application to the case of causation. This helps us move beyond the simple observation that the concept of causation seems to be disunified in complex ways, to an understanding of why this situation came about and how it is sustained.

FURTHER READING

Anscombe (1971) is a path-breaking discussion that has inspired a lot of subsequent work. Hall (2004) is a clear presentation of a simple form of pluralism that locates the view in relation to mainstream thinking about causation. (The collection in which Hall's paper appears, *Causation and Counterfactuals* (Collins, Hall, and Paul 2004), is a useful resource containing other relevant work.) The tension between difference-making and production as causal relations is also addressed in Hitchcock (1998) and Schaffer (2000). Within philosophy of science, Skyrms (1984) and Cartwright (1999) are good examples of pluralist work.

REFERENCES

ANSCOMBE, G. E. M. (1971). 'Causality and Determination,' repr. in E. Sosa and M. Tooley (eds.), *Causation*. Oxford: Oxford University Press, 1993: 88–104.

BEEBEE, H. (2004). 'Causing and Nothingness', in Collins, Hall, and Paul (2004: 291–308).

BRANDOM, R. (2000). *Articulating Reasons: An Introduction to Inferentialism*. Cambridge Mass.: Harvard University Press.

CARTWRIGHT, N.(1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

—— (2004). 'Causation: One Word Many Things', *Philosophy of Science* 71: 805–19.

COLLINS, J., HALL, E., and PAUL, L. (eds.) (2004). *Causation and Counterfactuals*. Cambridge Mass.: MIT.

GALLIE, W.B.(1956). 'Essentially Contested Concepts', *Proceedings of the Aristotelian Society* 56: 167–98.

GOPNIK, A., and SCHULZ, L. (2004). 'Mechanisms of Theory-Formation in Young Children', *Trends in Cognitive Sciences* 8: 371–7.

HALL, E. (2004). 'Two Concepts of Causation', in Collins, Hall, and Paul (2004: 225–76).

HEALEY, R. (1994). 'Nonseparable Processes and Causal Explanation', *Studies in History and Philosophy of Science* 25: 337–74.

HITCHCOCK, C. (1998). 'Causal Knowledge: That Great Guide of Human Life', *Communication and Cognition* 31: 271–96.

—— (2003). 'Of Humean Bondage', *British Journal for the Philosophy of Science* 54: 1–25.

HORWICH, C. (1990). *Truth*. Oxford: Clarendon.

LEWIS, D. K. (1973). 'Causation', *Journal of Philosophy* 70: 556–67.

LONGWORTH, F. (forthcoming). 'Is Causation a Cluster Concept?'.

SCHAFFER, J. (2000). 'Causation by Disconnection', *Philosophy of Science* 67: 285–300.

SKYRMS, B.(1984) 'EPR: Lessons for Metaphysics', in P. French, T. Uehling, and H. Wettstein, *Midwest Studies in Philosophy*. Minneapolis: University of Minnesota Press, ix. 245–55.

SOBER, E.(1985). 'Two Concepts of Cause', in P. Asquith and P. Kitcher (eds.), *PSA* 1984. East Lansing: Philosophy of Science Association, 405–24.

PART IV
THE METAPHYSICS OF CAUSATION

CHAPTER 17
PLATITUDES AND COUNTEREXAMPLES

PETER MENZIES

1. INTRODUCTION

One familiar method for studying the concept of singular or token causation is to consider the platitudes that the folk take for granted about the subject. Sometimes this method is associated with the view that the folk possess a tacit theory of causation in much the same way that they are commonly assumed to possess a tacit theory of grammar or a tacit theory of mind. On this view, the 'folk theory of causation' consists of a number of platitudes or principles that are taken to be common knowledge among ordinary subjects. These principles, or at least a substantial number of them, are assumed to be analytically true of the concept of causation. (For example, see Menzies 1996; Norton 2007.) Some philosophers who accept this view claim that the concept of causation can be defined in terms of these platitudes, or more generally in terms of their role in the folk theory, employing one of the standard treatments for defining theoretical concepts (as described, for example, in Lewis 1970). It is not necessary, however, to hold this view to think that the platitudes ordinary people associate with causation should play an important role in philosophical discussions about causation. Even philosophers who dispute that these platitudes are analytically true of causation think that it is important to pay attention to those beliefs about causation that are widely shared among ordinary subjects. For they think that such platitudes should act as constraints on any systematic theory of causation, whether reductive or not, in the sense that any such theory must be able to recover a fair number of the platitudes, or at the very least explain their seeming plausibility.

So it is not surprising that many philosophers appeal to the method of platitudes in their discussions of causation. What is more surprising is that there seems to be little *explicit* agreement among philosophers about which are the crucial platitudes about token causation. Consider the following small sample of philosophers who appeal to platitudes either as central principles of a folk theory of causation or as commonly held beliefs used to motivate particular analyses of the concept. In David Lewis's original formulation (1973) of his counterfactual theory of causation, he appeals to the intuition that causation is a transitive, difference-making relation to motivate his particular version of the counterfactual theory. D. H. Mellor (1995) deploys various 'connotations of causation' as constraints on acceptable analyses of causation, where the connotations are that causes explain and predict their effects and that causes can be used as means to achieving their effects. John Norton (2007) argues that the crucial platitude of the folk theory of causation is that causes bring about or produce their

effects and that causal processes are processes of production. Menzies (1996), on the other hand, argues that the crucial platitude is that causation is the intrinsic relation that typically accompanies the relation that holds when one event increases the probability of a distinct event.

Nonetheless, I suggest that this seeming diversity of views about the crucial platitudes is misleading. For behind the diversity of views there is a *hidden consensus* about one crucial platitude about causation. This belief about causation is presupposed by many of the other commonly cited platitudes, or taken as a given by almost all philosophical theories of causation. This is the platitude that, roughly speaking, *causation is a natural relation between events*. Broadly speaking, natural relations are external relations such as temporal and spatial distance that play an important role in the scientific conception of reality; and events are particular occurrences in time and space, not necessarily restricted to changes. So understood, this platitude forms the background to much philosophical theorizing about causation, but it is so much in the background that it is rarely the subject of explicit discussion.

In this chapter I wish to subject this philosophical platitude to critical scrutiny to see whether it is actually part of the folk conception of causation. I shall contend that its almost universal acceptance among philosophers is at variance with the evidence of the everyday and scientific use of the causal concept. Its inconsistencies with the use of the causal concept are so obvious that it is surprising that the platitude commands such agreement among philosophers. In contradiction to the platitude, I shall argue that that the concept of causation has essential elements that are contrastive, normative, and context-sensitive in character, and that these elements do not seat easily with the conception of causation as a natural relation. The moral I shall draw is that philosophical discussions need to attend to these elements of the causal concept in framing platitudes to constrain or motivate systematic theories of causation. This is not a rejection of the method of appealing to platitudes, but rather a plea to reject the incorporation of unquestioned metaphysical assumptions into this method. If we are to make progress in understanding the true nature of the causal concept we need to take care that such unquestioned metaphysical assumptions do not distort our understanding of it.

My plan of action is as follows: in sect. 2, I explain the conception of causation as a natural relation in more detail. In sect. 3, I outline some of the features of our use of the causal concept that do not fit with the idea of causation as a natural relation between events. In sects. 4 and 5, I outline the correct explanation of these features, replacing the metaphysical conception of causation with a conception of causation in terms of a contrastive difference-making relation, where the contrasts are determined contextually on the basis of what are often normative considerations.

2. Causation as a Natural Relation

The thesis that causation is a natural relation between events is explicitly formulated in P. F. Strawson's influential article 'Causation and Explanation'. Here Strawson (1992: 109) draws a distinction between causation and explanation in these terms:

We sometimes presume, or are said to presume, that causality is a natural relation which

holds in the natural world between particular events or circumstances, just as the relation of temporal succession does or that of spatial proximity. We also, and rightly, associate causality with explanation. But if causality is a relation which holds in the natural world, explanation is a different matter. ... [W]e also speak of one thing explaining, or being the explanation, of another thing, as if explaining was a relation in the sense in which we perhaps think of causality as a natural relation. It is an intellectual or rational or intensional relation. It does not hold between things in the natural world, things to which we can assign places and times in nature. It holds between facts or truths.

Strawson expands on the points about explanation later (ibid. 111):

As a first approximation, one could say that the non-natural fact that the explaining relation holds between the fact that p and the fact that q expands into the natural fact that coming to know p will tend, in the light of other knowledge (or theory) to induce a state which we call 'understanding why q'. The non-natural relation between the truths is mediated by the connection which, as matter of natural fact, we give them (or they have) in our minds. This is why, as a variant on calling the relation non-natural, I called it rational.

In these passages Strawson emphasizes two features of the distinction between causation as natural relation and explanation as non-natural or rational relation. First, causation is an extensional relation that connects real occurrences in the world whereas explanation is an intensional relation that connects facts or truths. Secondly, the existence of a causal relation does not depend in any way on the existence of human minds or the mental processes they engage in, whereas the existence of an explanatory relation does. Unfortunately, Strawson does not give a direct characterization of a natural relation.

What more can be said about the kind of natural relation that causation is presumed to be? There is no universally accepted catalogue of the features that a relation must have in order to count as a natural relation. Nonetheless, I suggest that there is a cluster of features that are typically associated with natural relations in philosophical discussions. First, as Strawson emphasizes, *natural relations relate concrete occurrences* that have temporal and spatial locations. We can adopt his terminology and call these 'events', so long as they are conceived broadly to include static as well as dynamic states of affairs. Secondly, natural relations are *contingent relations that are known a posteriori*. In other words, they are relations that may or may not hold in the actual world; and whether or not they do so is typically known on the basis of the usual scientific methods of observation and inference. The paradigm examples of such relations, sometimes called external relations, are temporal succession and spatial distance. Thirdly, *natural relations are non-evaluative and non-normative*. This is to say that they are the kind of relation that holds independently of any normative or evaluative judgements made by human subjects. A proposition about a natural relation in this sense does not imply, nor is it implied by, any proposition about norms or values. Fourthly, they are *relations that belong to a scientific natural kind*. This is somewhat vague, because what constitutes a scientific natural kind is somewhat indeterminate. But roughly the idea is that a scientific natural kind includes properties and relations that 'carve nature at its joints' in the sense of picking out objective

similarities in the world that form a reliable basis for scientific inference.

These are the features that philosophers typically associate with the idea of a natural relation. It is worth remarking that several philosophers have proposed a more specific, more metaphysically laden conception. For example, D. M. Armstrong's theory of universals (1978) and David Lewis's (1983) theory of perfectly natural properties and relations present more detailed conceptions that go beyond the features listed above. Although they differ in their conception of properties and relations, they agree on the extra duties that natural properties and relations have to perform. Thus while they suppose that natural properties and relations must carve nature at its joints, they argue that only those properties and relations that are *perfectly natural* can do this with complete success. Furthermore, they suppose that these perfectly natural properties and relations, or universals as Armstrong calls them, are very sparse in number: there are only as many as is required to characterize the way things are in reality comprehensively and without redundancy. Moreover, they both suppose that physics provides us with an inventory of the perfectly natural properties and relations that are instantiated in the actual world. Finally, Armstrong and Lewis suppose that all the perfectly natural properties and relations are intrinsic.

Whatever the metaphysical merits of Armstrong's and Lewis's conceptions of natural properties and relations, we do not need them for the purposes of our discussion. For our purposes it is necessary only to think of natural properties and relations as carving nature at its joints in the sense of marking objective, scientifically relevant similarities and differences among objects. We need to be able to say that green has more chance of being a natural property than grue; and that the relation of spatial distance is a better candidate for a natural relation than the relation of being collocated within the same suburb. To accept that there is some useful distinction between these kinds of property and relation does not require the heavy-duty metaphysics that Armstrong and Lewis deploy. We do not need to buy into their assumption that there is an elite class of perfectly natural properties and relations; or their assumption that physics tells us which members of this class are instantiated in the actual world, or even their assumption that the members of this class can serve as the basis on which the complete qualitative character of everything there is, and everything there could be, supervenes. Indeed, the purposes of this chapter require the more modest conception of natural relations. For I shall go on to dispute the widely held view that causation is a natural relation; and so my arguments will be more cogent if they are directed against a view that builds in as few contentious metaphysical assumptions as possible.

3. THE EVIDENCE AGAINST CAUSATION AS A NATURAL RELATION BETWEEN EVENTS

In this section I adduce evidence to show that we do not intuitively conceive of causation as a natural relation between events. I believe there is plenty of evidence to show this, but I shall gather evidence under just four representative headings. Other philosophers (Ehring Ch. 19 below; Hitchcock 1996*a*; 1996*b*; 2003; Maslen 2004; Schaffer 2006) have noted the same features of our intuitive conception of causation but have drawn somewhat different conclusions about it.

3.1 Events

If causation is a natural relation between events, then this relation should hold no matter how the events are described. In other words, causal propositions should be extensional in the sense that the substitution of one event nominal for a second coreferential event nominal in a causal proposition should preserve the truth-value of the proposition. The reasoning here is straightforward: if the causal relation is a natural relation in Strawson's sense, then whether it holds between a pair of events should be a completely mind-independent matter, and so should not depend on how the events are referred to. Of course, to make good this claim one has to have an adequate understanding of what events are. But it has been remarkably difficult to find a satisfactory conception of events that vindicates the extensionality of causal propositions.

Before arguing for this claim, I note a feature of causal propositions that should appear anomalous from the perspective of causal naturalism. This is the feature that many causal propositions cite particular objects rather than events as causes and effects. For example, it would not be unusual to come across a causal proposition to the effect that a particular gene caused a certain polypeptide molecule. The conventional response is to say that such propositions are elliptical for more complete propositions that mention events that involve the corresponding objects. So, for example, the proposition above is elliptical for a proposition such as the functioning of the particular gene caused the synthesis of the polypeptide molecule. The non-elliptical canonical causal proposition is always one that has events in the cause-and-effect places. Whatever the intrinsic plausibility of this view, I simply note that a unified treatment that handles all the different ostensible causal relata, including objects, in the same way would have the virtue of being simpler and less ad hoc.

Returning to the main issue, how should we conceive of the events that are the relata of the natural causal relation? Traditionally, events so understood are said to be happenings or occurrences that are as much real features of the world as physical objects, though they are categorically different. They are parasitic on physical objects in the sense that they would not exist if there were no physical objects. Their dependence on physical objects supplies the only sense in which they have spatial location: their spatial location is that of the physical object on which they depend. On the other hand, they have a temporal location in their own right: the temporal location is the time at which they occur. Beyond these features traditionally associated with events, some philosophers have championed more specific understandings of events. Davidson (1967; 1969), for example, argued for an understanding of the events as concrete particulars that are supposed to be analogous in certain fundamental ways to physical objects. This analogy forms the basis of two points he makes about events. First, just as definite description standing for a physical object need not provide a complete description of the object, a definite description standing for an event need not provide a complete description of the event. Thus, the nominal 'Smith's fall from the rockface' may pick out the whole cause of Smith's death even though it does not describe every detail of the cause. Secondly, just as there may be several definite descriptions standing for the same physical object, so there may be several definite descriptions standing for the same event. To take one of his examples, 'my flipping the switch', 'my turning on the light', 'my illuminating the room', and 'my alerting the burglar' do not refer to different events, but rather to one event described in different

ways.

In Davidson's view, it is events, so conceived of as concrete particulars, that serve as causation's relata. However, *pace* Davidson, they do not seem well suited to act in this role. The entities that are causation's relata are more finely individuated than concrete particulars. While Davidson supposes that 'my flipping the switch' and 'my alerting the burglar' provide different descriptions of the same event, ordinary causal discourse does not treat them as describing the same causal relatum. For replacing one description with the other in a causal proposition can change its truth-value. We may judge, for example, that flipping the switch caused electrons to race along the wire to the light bulb, but my alerting the burglar did not. Conversely, we might judge that the burglar being in the front room of the house when I entered was a causal condition of my alerting him, but it was not a causal condition of my flipping the switch. Of course, Davidson might retract his judgement about his example, saying that these descriptions refer, after all, to different descriptions. However, the same problem arises with other examples in which event nominals, uncontroversially referring to the same event, fail to be substitutable *salva veritate* in causal propositions. For example, to cite an example of Goldman (1970), the nominals 'my saying "hello"' and 'my saying "hello" loudly' are commonly thought to refer to one and the same concrete particular event, but the nominals are not intersubstitutable in causal propositions: my saying 'hello' loudly caused my neighbour to be startled, but saying 'hello' did not; and conversely, my nervousness about greeting my neighbour caused me to say 'hello' loudly, but did not cause me to say 'hello'.

Examples of this kind have convinced a number of philosophers (Goldman 1970; Kim 1969; 1973) to adopt a more fine-grained conception of events that are causation's relata. According to these philosophers, an event is to be thought of as a structured complex: complex because it consists of a number of different entities and structured because these entities must be put together in the right way to constitute an event. The constituents of such events are typically identified as a property F (or more generally an n-adic relation), a physical object x (or more generally an n-tuple of physical objects) and a period of time t (including both instants and intervals of time). To constitute an event these constituents must be related in a certain way. An event, which is represented symbolically as $[F, x, t]$, consists in the object x exemplifying the property F in the period of time t. The relation of an object exemplifying a property at a time is a primitive, unanalysable relation. Now under any plausible conception of properties, the property of alerting the burglar and the property of flipping the switch are different, as indeed are the properties of saying 'hello' and saying 'hello' loudly; hence the corresponding pairs of events are different as well. So the fact that substituting one event for another in a causal proposition does not preserve truth-value does not impugn the extensionality of causal propositions.

The strategy of making events to be more fine-grained than Davidson's concrete particulars is successful to some extent in saving extensionality. However, it does not go as far as it needs to go. For there are other examples that seem to indicate that causal propositions fail to be extensional even when events are individuated in a fine-grained way. A number of philosophers (e.g. Dretske 1977; Sanford 1985) have argued that we must include event aspects in addition to events as causal relata. They cite examples such as 'The height of the climber's fall caused his death' and 'The climber's falling *from a height* caused his death' in support of this claim. Achinstein (1983: ch. 6) has taken this observation further, arguing that

causal propositions are not always extensional. The first step in his argument is the assumption that nominals differing only in which of their elements is given contrastive stress refer to the same fine-grained event, so that the nominal 'Socrates' *drinking hemlock* at dusk' refers to the same event as the nominal 'Socrates' drinking hemlock *at dusk*'. The second step of the argument is the claim that these nominals, though co-referential, cannot be substituted *salva veritate* in the same causal propositions. For example, while it may be true that Socrates died because he *drank hemlock* at dusk, it is not true that he died because he drank hemlock *at dusk*. The explicit conclusion Achinstein draws is that causation is not an extensional relation after all.

Not too surprisingly, defenders of fine-grained conceptions of events have seen this kind of counterexample to extensionality as less than conclusive. As remarked earlier, some philosophers such as Dretske and Sanford have posited event aspects as causal relata and their response to Achinstein's argument might be to say that the extensionality of causal propositions requires only that the substitution of nominals referring to the same events or same event aspects preserve truth-value. But event nominals with different contrastive emphases denote different event aspects. David Lewis (1986*c*) adopts a different response, embracing a fine-grained conception of events that is different from Kim's. On his conception, an event is to be identified with a set of actual and possible spacetime regions, intuitively those in which the event does or could take place. So, for Lewis as for Kim, my saying 'hello' and my saying 'hello' loudly are different events, as are the events of my flipping the switch and my alerting the burglar. Lewis, however, goes beyond Kim in appealing to the idea that events have essences and accidental properties. This is useful precisely because it enables Lewis to distinguish the events of Socrates' *drinking hemlock* at dusk and Socrates' drinking hemlock *at dusk* on the grounds that they have different essential and accidental properties. The first event is essentially a drinking of hemlock by Socrates and only accidentally something that takes place at dusk, whereas the opposite is true of the second event, which is essentially something that happens at dusk and only accidentally a drinking of hemlock by Socrates. But if it is correct to think that the function of contrastive stress is to highlight the essential properties of events and that events with different essential properties are distinct, then Achinstein's example does not represent a genuine counterexample to the extensionality of causation. For the event nominals with their different contrastive emphases denote quite different events.

Whichever of these strategies is invoked to save the extensionality of causal propositions, it incurs some explanatory obligations. The strategies that appeal to event aspects or event essences have to explain what they are. I do not have decisive reasons for rejecting either of these strategies. However, I would argue that once one sees how it is possible to give a uniform and explanatorily parsimonious account of all the different kinds of things that can be said to be causes or effects, then one will find the ontological extravagances of event aspects and event essences too costly.

3.2 The Fragility of Events

If causation is a natural relation, then the events that are its relata should have determinate, mind-independent identity conditions. In particular, we should be able to answer questions

about their identity conditions, such as: Is it possible for an event that occurs at a certain time and in a certain manner to occur at a different time or in a different manner? This question is sometimes expressed in terms of terminology introduced by David Lewis (1986*b*: 196). He says an event is *fragile* to the extent that it could not have occurred at a different time or in a different manner. A fragile event is one, he says, that has a rich essence with very strict conditions for its occurrence. In terms of this terminology, the above question becomes: if causation is a natural relation between events, does it relate fragile events or unfragile events? Perhaps linguistic indeterminacy militates against there being a completely determinate answer to this question. Nonetheless, it is a reasonable expectation that we should be able to specify, within these bounds of indeterminacy, the identity conditions of events that are supposed to be the causal relata; and we should be able to specify them in a mind-independent fashion. But it seems quite difficult to do this.

Some philosophers have thought that the most fitting way to understand events as causal relata is to understand them as all uniformly very fragile in character. The thought is that if causation is really a relation carved out as a joint in nature, then it surely relates particular events with specific essences. Some causal propositions do apparently describe causal relations between extremely fragile events. For example, in response to the question of why a person died with a neutrino passing through his body, we might reply that an explosion on the sun caused the man's death as it happened in that very particular way. But normally it would be inappropriate to cite the explosion on the sun as the cause of a person's death since a death is not conceived to be so fragile as to include every specific detail of its occurrence. For the most part the causal propositions considered in everyday life and in scientific practice concern events that are not especially fragile in character. Indeed, if we were to think that all causally related events are fragile, we would open the gate to a flood of spurious causes. Lewis (1986*b*: 198) gives this example: suppose a gentle soldier on a firing squad does not shoot when the other members of the squad fire. There is a minute difference made by firing eight bullets instead of nine, and so if we suppose that the victim's death is fragile, we have to conclude that the gentle soldier caused the death by not shooting. The victim's very specific death depended on the gentle soldier's omission as much as the actions of the other members of the firing squad. I agree with Lewis that this result is a *reductio* of the strategy of treating effects as very fragile.

A natural response to this objection would be to deny that causal naturalism is necessarily committed to the idea that all events must be construed as extremely fragile. There is nothing inconsistent, after all, in supposing that natural causal relations relate not-so-fragile events. This may be true. What is inconsistent with the conception of causation as a natural relation is the fact that the fragility of events is a highly context-sensitive matter. Again I use an example of Lewis's (1986*b*: 198) to illustrate the point. An assassin kills his victim with poisoned chocolates, which the victim consumes after eating a big dinner. The victim's consumption of the big dinner slightly affects the time and manner of his death because poison taken on a full stomach passes more slowly into the blood. If we suppose the death is extremely fragile, then we must say that one of its causes is the eating of the dinner. But whether it is reasonable to cite the eating of the big dinner as a cause of the death depends on context. Suppose the poisoner is horrified by the victim's lingering death and says 'If only he hadn't eaten, this

wouldn't have happened,' where 'this' refers to the very fragile death. In this context it is reasonable to think of the effect as very fragile. But in other contexts it is reasonable to employ more lenient standards of fragility and say that the eating of the dinner did not cause the death.

There are still more complicated examples of the context-sensitivity of the fragility of events. Several philosophers (e.g. Bennett 1987; Mackie 1992) have noted that there is a general asymmetry in the way we treat hasteners and delayers. We are much more likely to cite as a cause an event that hastens some effect than we are to cite an event that delays the same effect. For example, we are prepared to say that a doctor who has hastened a patient's death by injecting the patient with a lethal dose of morphine has caused the death; but not so prepared to say that a doctor who has delayed a patient's death by administering life-prolonging drugs has caused the patient's subsequent death. Despite this general asymmetry, there are special contexts in which it is reasonable to cite a delayer as a cause of some effect. For example, suppose that a doctor has administered a drug that means the patient dies several days later than he would normally have died. If we are interested in why the patient died later than he would normally have, it is quite appropriate to cite the doctor's intervention as the cause of the delayed death.

Examples of this kind suggest that the fragility of events is a context-sensitive matter, and so dependent on the way human minds determine the values of certain contextual parameters.

3.3 Absences as Causes and Effects

The third category of examples that appear to pose a problem for causal naturalism concerns causal propositions that cite absences, omissions, or lacks as causes or effects. For example, we assert causal propositions to the effect that lack of food causes one's hunger and that vaccination prevents one from catching a disease or, in other words, causes one not to catch a disease. Such causal propositions present a problem because such absences do not appear to be genuine events; and in so far as such absences can be causes and effects, causation cannot be a relation, natural or otherwise, between events. Where there are no relata, there can be no relation. (On this point see Mellor 1995: ch. 11; Lewis [2000] 2004.)

Of course, one might claim that, despite appearances to the contrary, absences are genuine events after all. One possible defence is that absences such as lack of food and absence of disease are very special events that occur whenever some other event does not occur: for example, a lack of food event occurs just when no ingestion of food occurs; and an absence of disease event occurs when no disease event occurs. But these special events are very mysterious indeed! Their existence would refute the plausible Humean claim that existential claims about distinct entities are logically independent. Another possible defence of absences as events is to claim that nominals such as 'lack of food' or 'absence of disease' refer instead to genuine positive events occurring at the same time and place. Perhaps, 'the lack of food' refers to the normal processes that go on in the digestive system when no food is ingested and 'the absence of disease' refers to the normal processes of cell functioning when disease is not present. However, this claim does not appear to be defensible. The causes and effects of a lack of food—famine on the one hand and hunger on the other—are very different from the causes

and effects of the metabolic processes that go on when food is not being ingested. Again, the causes and effects of a doctor's failure to administer care to a patient are very different from the causes and effects of the doctor's taking a nap during the relevant period of time.

Given the implausibility of this defence, it is hard to avoid the conclusion that absences should not be identified with positive occurrences. But some have denied that this shows that causation is not a natural relation between events. These philosophers have adopted a number of strategies. One strategy adopted by Beebee (2004) is to argue that genuine causal propositions describe causation as a relation between events while those propositions that appear to cite absences as causes or effects are really causal explanations. Such a proposition is really giving explanatory information about the network of positive causal relations. Another strategy adopted by Fair (1979), Dowe (2000), and Armstrong (2004) is to distinguish a primary class of genuine causal propositions that describe causal relations between events from a secondary class of propositions that cite absences as causes or effects; and then to argue that the propositions in the secondary class are parasitic or dependent on propositions in the first class. These philosophers attempt to make good the claim of dependence by analysing propositions about absences as causes and effects in terms of counterfactuals about causation between positive events. For example, sample counterfactual analyses might run along these lines: the absence of any event of kind C directly causes positive event e iff, if there had been a positive event c of kind C, c would or might have caused some positive event d incompatible with the event e; positive event c directly causes the absence of any event of kind E iff c causes some positive event d incompatible with any event of kind E. (See Lewis [2000] 2004: 284–5.)

There is much that can be criticized about the details of these counterfactual analyses. But a more general criticism of this second strategy, as well as of the first strategy, is that there is no psychological evidence that people draw a sharp distinction between causation involving positive events and causation involving absences. Indeed, both strategies assume that there is some hard and fast distinction between genuine events and absences. But there is considerable evidence to the contrary that the distinction between positive and negative occurrences is a context-sensitive one: what is taken to be an event or positive occurrence in one context can be taken to be an absence or negative occurrence in a different context. To take an example discussed by Woodward (2006), a person's death is often treated as an unproblematic event in commonplace contexts, especially in legal and moral contexts, where a death can be investigated in formal inquests as something that has causes (asphyxiation, starvation) and effects (bodily decomposition, grief). But now consider a somewhat unusual context in which we focus on the complex and intricate coordination of biochemical processes that sustain life and the death of an organism looks unsurprising. In this context, the life of an organism is taken to be a positive surprising occurrence and the death of an organism is viewed as an unsurprising part of the normal course of events. In this context, death is treated as an absence of life or life-sustaining processes. More generally, I agree with those philosophers (e.g. Hitchcock 2001; Maudlin 2004; Woodward 2006) who claim that the distinction between positive and negative occurrences reflects a deeper contrast between what might be described as the normal or default outcomes in some kind of situation on the one hand and departures or deviations from it on the other hand, with the former typically being called absences and the latter presences. On this understanding, the default/deviation contrast is highly contextual and

theory-relative. (These claims are supported by a great deal of psychological evidence: see Kahneman and Tversky 1982; Kahneman and Miller 1986; Byrne 2002.)

These considerations suggest that we should not seek to draw a hard and fast distinction between causation involving events and causation involving absences. Indeed, some see them as showing that causation is not, properly speaking, a relation at all. These philosophers claim, as Mellor (1995; 2004) does, that causal propositions really concern the way in which facts cause other facts, where these facts may be about the occurrence or non-occurrence of events. Or they claim, as Lewis (1973; 2004), does that causal propositions attempt to convey counterfactual truths about the dependence of one proposition on another, where these propositions may be about the occurrence or non-occurrence of events. While these philosophers argue that causation is not a natural *relation*, they would nonetheless maintain that causation is all the same a natural *phenomenon*. In response to this more subtle form of causal naturalism, I would press the objection that they cannot adequately distinguish between absences that are genuine causes and effects on the one hand and absences that are only spurious causes and effects on the other. As we have seen, there are cases where we genuinely want to say that an absence is a cause. For example, a doctor's failure to administer a life-saving drug to a patient is the cause of the patient's death. But the problem is that the counterfactual accounts of causation accepted by Mellor and Lewis render countless other absences as causes of the patient's death. The failure of the hospital cleaner to administer the drug is something without which the patient would not have died, or would have had a much lower chance of dying than otherwise. But we draw a clear difference in causal status between the doctor's omission and the hospital cleaner's omission. Moreover, there is ample evidence that normative considerations play a crucial role in determining the difference between these absences. For it was the doctor's, not the hospital cleaner's, *duty* to care for the patient and this fact implies that the doctor's omission was a cause of the death in a way that the hospital cleaner's omission was not. (For extensive discussions of the role of normative considerations in causal claims see Knobe and Fraser 2008; Hitchcock and Knobe forthcoming.) The fact that normative considerations play a crucial role in the explanation of the distinction between genuine and spurious causal absences suggests that the truth of causal propositions is not such a perfectly natural phenomenon as these defenders of causal naturalism would have us believe. But these considerations merge into considerations about the final category of problematic cases, to which we now turn.

3.4 Causes vs. Background Conditions

The fourth and final category of examples that pose problems for the conception of causation as a natural relation, or at least a natural phenomenon, concerns the distinction we intuitively draw between causes and background conditions. The problem is that this distinction seems to be context-sensitive and theory-relative in ways that are inconsistent with the idea that the truth-conditions of causal propositions are a completely objective, mind-independent matter.

It is commonly acknowledged that the truth or falsity of causal propositions is relative to context. One kind of context-relativity, sometimes called relativity to the context of

occurrence (Gorovitz 1965), does not seem so problematic. If a building is destroyed by fire, the main cause might be cited as the sparks set off by faulty electrical wiring. It is also true that the fire would not have taken hold but for the oxygen in the air, the presence of combustible material, and the dryness of the building. But these are mere background conditions of the fire. On the other hand, if a fire breaks out in laboratory or in a factory, where special precautions are taken to exclude oxygen during the experiment or manufacturing process, it would be appropriate to cite the presence of the oxygen as a cause of the fire. This kind of relativity is not problematic because the different contexts involve objective structures that can ground the different causal judgements.

A more problematic kind of context-relativity for causal naturalism is known as relativity to the context of enquiry (ibid.). Rather than two situations eliciting different judgements about causes and conditions, this kind of case involves different judgements being made about the same situation depending on the type of enquiry being undertaken. Hart and Honoré ([1959] 1985: 35–8) give some examples of this kind of case. The cause of the great famine in India might be identified by an Indian peasant as the drought that preceded it, but the World Food Authority might identify the Indian government's failure to build up food reserves as the cause and the drought as mere condition. Or to take a different but familiar example: someone who prepares meals for a person suffering from an ulcerated condition of the stomach might identify spicy food as the cause of indigestion, but a doctor might identify the stomach ulcers as the cause of the indigestion and the meal as a mere condition. In this kind of case the context of enquiry determines what counts as cause and what as background condition. This is problematic for the view that causation is a natural relation, or at least a natural phenomenon, because it indicates that the truth-conditions of causal propositions must take into account the way human subjects enquire about causation.

The relativity of causal propositions to a context of enquiry has been often acknowledged and discussed as the phenomenon of causal selectivity, of how 'the cause' is selected from a set of objective causal conditions for some event. But philosophers from Mill to Lewis have claimed that the principles of selection are capricious or invidious. They maintain that a broad non-discriminatory notion of 'a cause' is the primary causal notion so that there is a set of objective causes satisfying mind-independent conditions for any given effect and that pragmatic principles of selection operate to pick out a salient factor to be labelled 'the cause' from this set of objective conditions.

This whole approach is highly questionable in my view. In the first place, the way we distinguish between cause and condition is not capricious or invidious. There is often widespread agreement among speakers about how the distinction should be drawn in any case. There must be some basis to this systematic agreement that deserves theoretical elucidation. Secondly, the distinction is not just a minor, peripheral feature of causal usage, but is actually central and crucial to the concept of causation. As Hart and Honoré remark: 'The contrast between cause and condition is an inseparable feature of all causal thinking, and constitutes as much the meaning of causal expressions as the implicit reference to generalizations does' ([1959] 1985: 12). Thirdly, it is a philosophers' myth that the notion of 'a cause' is the primary causal notion of common-sense and scientific usage; and equally a myth that this is a completely objective notion displaying none of the context-relativity of the notion of 'the cause'. The expression of 'the cause of the explosion' does not usually mean 'the most salient

among the factors that count as a cause of the explosion'. Rather the direction of explanation is the other way: the expression 'a cause of the explosion' is understood as meaning 'one of the causes of the explosion' with the expression 'a cause' inheriting all the selectivity of the expression 'the cause(s)'.

4. CAUSES AS DIFFERENCE-MAKERS FROM THE NORMAL COURSE OF EVENTS

How are we to explain the phenomena that appear to cast doubt on the idea that causation is a natural relation between events? In this section I sketch a theory of causation that straightforwardly explains these phenomena.

The starting point for the theory is the idea that causation is intimately connected with the idea of making a difference: a cause is something that makes a difference to its effects. Many philosophers of causation appeal to this idea to motivate their theories but formulate it in slightly different ways. The formulation that I find most insightful is one given originally by Hart and Honoré ([1959] 1985) in their classic work *Causation in the Law*:

Human action in the simple cases, where we produce some desired effect by the manipulation of an object in our environment, is an interference in the natural course of events which *makes a difference* in the way these develop. In an almost literal sense, such an interference by human action is an intervention or intrusion of one kind of thing upon a distinct kind of thing. Common experience teaches us that, left to themselves, the things we manipulate, since they have a 'nature' or characteristic way of behaving, would persist in states or exhibit changes different from those which we have learnt to bring about in them by our manipulation. The notion, that a cause is essentially something which interferes with or intervenes in the course of events which would normally take place, is central to the commonsense concept of cause, and at least as essential as the notions of invariable or constant sequence so much stressed by Mill and Hume. Analogies with the interference by human beings with the natural course of events in part control, even in cases where there is literally no human intervention, what is identified as the cause of some occurrence; the cause, though not a literal intervention, is a *difference* to the normal course which accounts for the difference in the outcome. (ibid. 29)

The passage is clear about what Hart and Honoré take to be the central guiding idea about causation: a cause is an intervention, analogous to a human action, that brings about changes in the normal course of events. I believe that this guiding idea has much immediate intuitive appeal, but its ultimate justification is the success it has in explaining a vast range of seemingly disparate and unusual features of the causal concept. This guiding idea nonetheless requires further elucidation to make its implications more transparent.

Fortunately, a well-developed theory of causation is at hand that can provide the basis for an explication of Hart and Honoré's guiding idea. I have in mind the theory of causation

developed by James Woodward (2003) in his book *Making Things Happen*, building on the pioneering work of Judea Pearl (2000) and Peter Spirtes, Clark Glymour, and Richard Scheines (1993). Woodward develops an account of causation at both the type and token levels, but I shall start by expounding his theory of type causation, later adapting it to provide an account of token causation. Like Hart and Honoré, Woodward believes that type causation is linked to the idea that a cause makes a difference to its effects. But he explicates this idea in terms of a framework that takes the relata of causation to be variables. He says that a causal relation obtains when a change in the variable representing the cause is associated with a change in the variable representing the effect. The variables involved may be many-valued or binary-valued. In simple cases of type causation the variables are binary variables the values of which indicate the presence or absence of a property. More precisely, Woodward's truth-condition for type causation is:

(1) Variable *X causes* variable *Y* if and only if, were an intervention to change the value of the *X* variable, it would change the value of the *Y* variable.

(This is a simplified version of Woodward's account of type causation that focuses on his definition of a *total* cause. He defines a *direct* cause differently. But in the simple cases we shall consider the definitions coincide.) This truth-condition implies that there must be at least two values of *X*—*x* and *x′*—and two values of *Y*—*y* and y′—such that if an intervention were to change the value of *X* from *x′* to *x*, the value of *Y* would change from *y′* to *y*.

Woodward's account, like Hart and Honoré's, emphasizes that the changes in the causally related variables must occur by virtue of an intervention on the cause variable, although the intervention may be hypothetical rather than actual. The reason for this emphasis is that spurious correlations can occur where changes in one variable are accidentally associated with changes in another without any causal relation between them. For example, decreases in barometer readings are correlated with onsets of storms, but this correlation is due to these phenomena being the effects of a common cause—drops in atmospheric pressure. However, if the changes in the barometer reading were brought about by an intervention, say by an experimenter fixing the reading of the barometer, the correlation with the onset of a storm would disappear. This is the central difference between correlations and causal relations: a genuine causal relation is robust under interventions that change the values of the cause variable.

However, Woodward's approach differs from Hart and Honoré's quasi-manipulability approach in that it does not link the existence of a causal relation with the possibility of a *human intervention*. The crucial difference lies in Woodward's definition of an intervention. For Woodward, an *intervention* on one variable *X* with respect to another variable *Y* is, roughly, an idealized experimental manipulation that causes *X* to change its value in such a way that all other variables that previously were causally relevant to *X* no longer influence it; and in such a way that any change in *Y* can only come about through the change in *X*. An intervention could be the result not only of a human action, but also of a 'natural experiment'. By generalizing the notion of an intervention and defining it in terms of explicit causal notions, Woodward's account eschews a reductive approach to causation. Nonetheless, in spite

of its non-reductive character, we shall see that the theory has non-trivial and illuminating implications about the structure of causal concepts and their interrelationships.

As remarked above, Woodward offers an account of token causation as well as type causation. Applied to token causation, Woodward says, the theory transposes, roughly speaking, from one that talks of causation relating *variables* to one that talks of causation relating *values of variables*. With token causal claims, the cause and effect are represented as the actual values of variables; that is, the values of the variables that happen to be realized in the actual circumstances. The content of a token causal claim is that an intervention that sets the cause variable at its actual value would thereby set the effect variable at its actual value. More precisely, the truth condition for token causation is stated:

> (2) $X = x$ *causes* variable $Y = y$ if and only if (i) $X = x$ and $Y = y$ are the actual values of these variables; and (ii) there are values of the variables X and Y—say, x' and y'—such that if an intervention were to change the value of the X variable from x' to x, value of the Y variable would change from y' to y.

To illustrate this truth-condition, suppose that Joe smokes two packets of cigarettes a day and gets lung cancer; and we represent these states as the actual values of the many-valued variable X that takes the values {Joe smokes no cigarettes a day, Joe smokes a packet of cigarettes a day, Joe smokes two packets of cigarettes a day, Joe smokes three or more packets of cigarettes a day} and the binary variable Y that takes the values {Joe gets lung cancer, Joe does not get lung cancer}. Then his smoking two packets of cigarettes a day caused his lung cancer because there are values of the variables—Joe's smoking no cigarettes a day and his not getting lung cancer—such that if an intervention were to change him from a non-smoker to a two-packets-a-day smoker, it would also change his not getting lung cancer to his getting lung cancer.

How well does Woodward's interventionist theory capture Hart and Honoré's guiding idea? It captures reasonably well the concept of cause as an intervention that makes a difference. However, it does not capture so well the idea that a cause is an intervention that makes a difference *to the normal course of events*. There is no mention of the notion of normality or normal course of events in Woodward's account of either type or token causation. But this is a central feature of Hart and Honoré's guiding idea. In order to capture this feature, it is necessary to introduce a distinction that has come to be called the distinction between *the default* and *deviant values* of variables. (See Hitchcock 2001; 2007; Woodward 2003.) The default value of a variable is the value that represents the state of a system that requires no explanation because it is normal or to be expected. On the other hand, a deviant value of a variable represents a state of a system requiring explanation because it is abnormal or anomalous in some sense. As we shall see, the precise understanding of these notions is sensitive to subtle and variable contextual cues. But as a preliminary elucidation, one can say there are at least three specific ways in which the distinction between the default and deviant values of a variable can be understood within a given context. (For a more detailed discussion see Menzies 2007; Hitchcock and Knobe forthcoming.)

First, the distinction between the default and deviant values of a variable sometimes reflects

the distinction between *statistical mean* value and the non-mean values of the variable. Thus, if the height of a particular soldier is different from the average height of soldiers in general, his height might be regarded as a deviant value of the variable height. Or if every planet in the solar system has been observed to conform to an elliptical orbit predicted by Newton's laws, then a planet deviating from an elliptical orbit represents an anomalous phenomenon that might be represented as a deviant value of the relevant variable.

Secondly, the distinction between default and deviant values sometimes reflects the distinction between the state of *a properly functioning* system and the state of the system when it malfunctions. To take a medical example, whether a human body is healthy or diseased is determined on the basis of the proper functioning of the body's various organs. Thus, a healthy state of a body is naturally represented as a default value of a variable representing health status and a diseased or unhealthy state as a deviant value. Again, the design of a computer system determines whether it functions properly or improperly. The distinction between the state of a properly functioning computer system and its state when it malfunctions can be represented in terms of the distinction between the default and deviant values of a variable.

Finally, the distinction between default and deviant values can represent the difference between *states that conform to some legal, moral, or social norm* and those states that do not so conform. For example, if a government has an obligation to set aside food reserves for its population, then its failure to do so might be represented as the deviant value of a variable. Or again, if it is regarded as impermissible to smoke in a public place, a person's refraining from smoking might be represented as the default value and the person's smoking as the deviant value of the variable representing their smoking status.

These are some of the ways in which the default/deviant value contrast is understood. One might wonder whether the specific ways in which I have claimed we understand the contrast really have much in common; and whether the general default/deviant contrast is of any theoretical utility. After all what is statistically normal is a very different from what is functionally normative and from what is legally, morally, or socially normative. This objection notwithstanding, I wish to claim that our reasoning about causation applies uniformly to all these different modes of understanding the default/deviant values contrast even though there are discernible differences between the modes. For there is indeed something in common to these modes, which is that the default values represent states of systems that are ideal in certain respects and deviant values represent states that are departures from these ideal states. The different modes of understanding default values of variables correspond to different dimensions or respects of idealness.

Finally, we are in a position to capture the last element in Hart and Honoré's guiding idea that a cause is an intervention that makes a difference *to the normal course of events*. In order to do this we need to introduce a contextual parameter that selects the default values of the cause-and-effect variable. The selected values represent what the normal course of events is for the system under consideration. Where the cause variable is X and the effect variable is Y, I stipulate that the default values of these variables are denoted by 'x'' and 'y'' respectively. I represent the contextual parameter with its default values set by context in terms of the ordered pair $X = x'$, $Y = y'>$. Accordingly, the truth-condition for token causation I propose is this:

(3) $X = x$ causes $Y = y$ relative to the context $X = x'$, $Y = y'>$ if and only if (i) the actual values of X and Y are x and y respectively; and (ii) if an intervention were to change the value of X from x' to x then the value of Y would change from y' to y.

(It is important to note that the situation in which cause and effect represent departures from the normal course of events is just one type of situation in which we attribute causation. There is another less common case in which the token cause and the effect are represented by the default values of the cause-and-effect variables and the contrast situations are represented by the deviant values. For further discussion of this kind of case, see Menzies 2004.)

The obvious difference between this truth-condition and Woodward's truth-condition for token causation is that the former relativizes the truth-condition to a contextual parameter determining default values whereas the latter deploys existential quantification over the values acting as contrasts to the actual values. It may be difficult to see the need to introduce the contextual parameter X ¼ x', Y ¼ $y'>$ if we focus on the case in which X and Y are binary variables. In such a case as soon as we know the actual values, which are stipulated to be deviant values, we can infer what the default values must be. But this need not be the case in general, as is most evident in the case of many-valued variables. Return to the example of Joe, a two-packet-a-day smoker who contracts lung cancer. On the proposed account of token causation, to order to determine whether Joe's smoking caused his lung cancer we have to determine the default values of the cause-and-effect variables. If a context deems it normal for Joe not to smoke at all and not to get lung cancer, then his smoking two packets a day counts as the cause of his smoking since an intervention that changed him from a non-smoker to a two-packet-a-day smoker would change his cancer status. On the other hand, if a context deems it normal for Joe to smoke three or more packets of cigarette a day, then his smoking two packets a day does not count as the cause of his lung cancer, since the corresponding intervention would not change his cancer status.

It will be convenient in future to have a canonical form for expressing token causal propositions, as they are understood in this framework. I shall use the form of expression: X's taking the value x rather x' causes Y to take the value y rather than y', where x' and y' will always be understood as representing the default values of the cause-and-effect variables. The virtue of this form of expression is that it brings to the surface the *contrastive character* of token causal claims. Let us see now how this framework might be invoked to explain the phenomena cited above as posing problems for causal naturalism.

5. EXPLANATION OF THE PHENOMENA

5.1 Events Explained

The problem under this heading was that the view of causation as a natural relation implies that causal propositions should be extensional. In an effort to preserve extensionality, adherents of the view are forced to embrace increasingly fine-grained conceptions of the relata of the causal relations, from Kimian structured complexes to event-aspects. This strategy is

ontologically costly in the way it posits many different kinds of causal relata and in the way it invokes onerous metaphysical distinctions between event essences and accidents. The strategy is unnecessary if we adopt the proposed contrastive conception of causation.

The proposed account has no trouble explaining how objects and events, as coarse-grained particulars, can be cited as causes and effects. To say that an object or event caused some other object or event is just a shorthand way of saying that the presence of the object or the event rather than its absence brought it about that another object or event was present rather than absent.

This proposed account also allows us to say that fine-grained events too can act as causes and effects. For example, my saying 'hello' loudly is a different cause from my saying 'hello', even though the occurrence of the first implies the occurrence of the second. The difference in causal status is explained in terms of the fact that the contrast relevant to my saying 'hello' loudly may be different from the contrast relevant to my saying 'hello'. In the case of my saying 'hello' loudly, the relevant variable might take any value in the set {I do not say 'hello', I say 'hello' softly, I say 'hello' loudly} with the second value as default, whereas in the case of my saying 'hello' the relevant variable might take values in the more restricted set {I do not say 'hello', I say 'hello'}, with the first value as default. Then it may be true that my saying 'hello' loudly rather than softly was responsible for my neighbour's being startled rather than not being startled, while it is false that my saying 'hello' rather than not saying 'hello' was responsible for this contrast.

Correspondingly, we can provide an explanation of the different causal roles of event nominals with different contrastive focus. Thus, 'Socrates' *drinking hemlock* at dusk' may have a different causal role from 'Socrates' drinking hemlock *at dusk*' even though these event nominals refer to the same coarse-grained event and perhaps indeed to the same fine-grained event. We do not have to posit event aspects, or indeed event essences and accidents, to explain this difference in causal role. The point of the contrastive focus in the event nominals is to indicate the range of values of the relevant variables. The relevant values in the case of 'Socrates' *drinking hemlock* at dusk' are those in the set {Socrates drinks hemlock at dusk, Socrates does not drink hemlock at dusk} while the relevant values in the case of 'Socrates' drinking hemlock *at dusk*' are those in the set {Socrates drinks hemlock at dusk, Socrates drinks hemlock at some other time}. Given this fact, it is not surprising that Socrates' dying rather than remaining alive is better explained in terms of the contrast between his drinking hemlock and not drinking hemlock rather than in terms of the contrast between his drinking hemlock at dusk and drinking hemlock at some other time.

This approach to the issue of causal relata makes no serious ontological commitment to any particular conception of the events, whether coarse- or fine-grained, or to event aspects. The question whether we should be committed to these entities must be settled on the basis of other considerations. In taking causation to relate contrasts, this approach only commits itself to the existence of the relevant contrasts; and this commitment is best understood simply as a commitment to the truth of a pair of counterfactuals. It is plausible that the content of the claim that 'X = x rather than X = x′ caused Y = y rather than Y = y′' is captured by the counterfactuals 'If it were the case that X = x′ then it would be the case that Y = y′' and 'If it were the case that X = x then it would be the case that Y = y'. (For discussion see Menzies

2004; 2007.)

5.2 The Fragility of Events Explained

The problem under this heading was that if causal naturalism is true it should be determinate within certain bounds whether the events related by causation are fragile or not. But the fragility of events seems to be a highly context-sensitive matter. In some contexts it is appropriate to cite a fragile cause and in other contexts not, making the issue of fragility turn on the way human minds fix the value of a contextual parameter.

The present account explains the fragility of events in terms of the fact that causes and effects are contrasts between the values of variables, with context determining the appropriate variables and their values. In particular, the degree of fragility of an event depends on the way context determines the contrast between values of *the effect variable*. So, for example, if the contrast to be explained in a context is why the firing squad victim died rather than did not die, it would not be appropriate to cite the contrast between the gentle soldier's not firing rather than firing. For an intervention that made the gentle soldier not fire rather than fire would make no difference to whether the victim died. Similarly, if the contrast to be explained in the other example is between the poisoning victim's dying rather than not dying, it would not be relevant to cite the contrast between earlier consumption of a big dinner versus non-consumption. However, if the contrast at the effect end is the contrast between his dying a lingering, painful death and his dying a quick, painless death then it would be appropriate to explain this contrast in terms of his earlier consumption of a big dinner.

It is possible to explain in similar fashion some of the puzzling features of the causal concept, such as the asymmetry between hasteners and delayers. For example, if we are explaining why a man died-at-time-t rather than survived-at-time-t (for some contextually relevant t), it is often appropriate to cite a hastener but not a delayer. This asymmetry is rooted in a certain objective feature of the world: there exist times t such that a hastener makes the difference between a person's dying-at-time-t and surviving-at-time-t, but there exist no times t such that a delayer makes a comparable difference. On the other hand, if we are interested in why a person died-at-time-t rather than died-at-time-t' (for contextually relevant t and t'), then a hastener and a delayer stand on an equal footing. The former asymmetry between hasteners and delayers obscures the latter symmetry because we tend to interpret the question 'Why did the person die?' as 'Why did the person die-at-time-t rather than survive-at-time-t?' rather than 'Why did the person die-at-time-t rather than die-at-time-t'?' (This explanation of the asymmetry between hasteners and delayers is due to Hitchcock: see Schaffer 2006: n. 15.)

5.3 Absences as Causes and Effects Explained

The problem for causal naturalism under this heading was that absences are commonly cited as causes and effects, but since they are not, strictly speaking, anything, they cannot act as the relata of any relation. We saw that various defences of causal naturalism against this objection overlooked the context-sensitive character of the distinction between positive and negative occurrences, or failed to account for the role normative considerations play in

causally discriminating among absences.

In contrast, the present account provides a ready explanation of the causal role of absences. The important thing here is that absences are commonly cited as causes and effects in situations in which some customary procedure, practice, or routine has been developed to neutralize or counteract some harm. For example, the harmful effect of drought is regularly neutralized by government precautions in conserving water; disease is neutralized by inoculation; rain by the use of umbrellas. When some harm occurs in violation of the expectations set up by these practices or routines, the cause is said to be an omission or failure on the part of some agent to carry out the neutralizing procedures. In such cases the omission is a deviation from the normal course of events that accounts for a subsequent harm. So the lack of food in contrast to the normal availability of food explains why a person starves rather than has a full stomach. A failure to vaccinate in contrast to routine vaccination explains a child's contracting a disease rather than staying healthy.

The present account does not have to appeal to recondite metaphysical doctrines to explain causal talk about absences. The linguistic function of talk about absences is to highlight contrasts rather than to refer to special kinds of negative occurrences. Nor does the account have any problem accommodating absences as causes without generating spurious causal absences. For example, the distinction in causal status between the doctor's omission and the hospital cleaner's omission in our earlier example is easily explained in terms of the default values of the relevant causal variables. In the normal course of events it is the doctor's duty to administer the life-saving drug to his patient, not the hospital cleaner's duty. Accordingly, the contrast between the doctor's omission and the normal course of events explains why the patient died rather than survived, but there is no comparable explanatory contrast between the hospital cleaner's omission and the normal course of events. And as we have seen, it is not unusual for the determination of what counts as the normal course of events or the default values of the variables to be determined on the basis of normative considerations. This may violate the strictures of causal naturalism, but commonsense attributions of causation are steeped in normativity of all kinds.

5.4 Causes vs. Background Conditions Explained

The problem for the causal naturalist under this heading was that distinction between causes and background conditions seems to be context-sensitive but nonetheless systematic and robust in ways that defeat the causal naturalist's claim that the distinction is capricious or invidious.

The current proposal makes the causal naturalist's ad hoc manoeuvres unnecessary. For it stipulates that the difference between a cause and background conditions is the difference between a variable that is assigned a deviant value and variables that are assigned default values. Cases of relativity to context of occurrence pose no difficulty. In an ordinary context in which the sparks given off by faulty electrical equipment are the cause of a building fire and the presence of oxygen is a background condition, the corresponding variables are assigned deviant and default values respectively. Whereas in the special context of an

experimental or manufacturing process that takes precautions to exclude oxygen, the corresponding variables are given the reverse assignment of deviant and default values.

Similarly, there is no difficulty explaining cases of relativity to a context of enquiry. For the peasant, the drought is the cause of the famine and the government's failure to build up food reserves a background condition, a fact reflected in the assignment of deviant and default values to the corresponding variables. On the other hand, for the World Food Authority the government's failure to build up food reserves is the cause and the drought a background condition, which is reflected in the reverse assignment of deviant and default values to the variables.

To be sure, the explanations of the examples under this heading do not go very deep. For the concept of a default value of a variable is really just a formalization of the concept of a background condition that is part of the normal course of events. All the same, the distinction between cause and background condition falls neatly into place as part of the general conception of a cause as an intervention that makes the difference to the normal course of events.

FURTHER READING

P. F. Strawson (1992) is the classic exposition of the view that causation is a natural relation. Chapter 19 of this volume, 'Causal Relata', reviews many of the same data concerning causal relata covered under the headings 'Events', 'The Fragility of Events', and 'Absences as Causes and Effects' above but reaches very different conclusions from this chapter. Hitchcock (1996*a*; 1996*b*; 2001; 2007), Maslen (2004), Schaffer (2006), and Woodward (2003) offer contrastive accounts of causation, though they do not emphasize the role of context and of normative considerations in contrast selection. Knobe and Fraser (2008) and Hitchcock and Knobe (forthcoming) stress the role of normative considerations in attributions of causal responsibility.

REFERENCES

ACHINSTEIN, P. (1983). *The Nature of Explanation*. Oxford: Oxford University Press.

ARMSTRONG, D. M. (2004). 'Going through the Open Door Again: Counterfactual versus Singularist Theories of Causation', in Collins, Hall, and Paul (2004: 445–57).

BEEBEE, H. (2004). 'Causing and Nothingness', in Collins, Hall, and Paul (2004: 291–308).

BENNETT, J. (1987). 'Event Causation: The Counterfactual Analysis', in J. Tomberlin (ed.), *Philosophical Perspectives*, i. *Metaphysics*. Atascadero, Calif.: Ridgeview.

BYRNE, R. (2002). 'Mental Models and Counterfactual Thoughts about What Might Have Been', *Trends in Cognitive Science* 6: 426–31.

COLLINS, J., HALL, E., and PAUL, L. (eds.) (2004). *Causation and Counterfactuals*. Cambridge Mass.: MIT.

DAVIDSON, D. (1967). 'Causal Relations', in Davidson (1980: 105–21).

—— (1969). 'The Individuation of Events', in Davidson (1980: 163–80).

—— (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

DOWE, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

DRETSKE, F. (1977). 'Referring to Events', *Midwest Studies in Philosophy* 2: 90–9.

FAIR, D. (1979). 'Causation and the Flow of Energy', *Erkenntnis* 14: 219–50.

GOLDMAN, A. (1970). *A Theory of Human Action*. Princeton: Princeton University Press.

GOROVITZ, S. (1965). 'Causal Judgements and Causal Explanations', *Journal of Philosophy* 62: 695–711.

HART, H., and HONORÉ, A. ([1959]1985). *Causation in the Law*. Oxford: Oxford University Press.

HITCHCOCK, C. (1996*a*). 'Farewell to Binary Causation', *Canadian Journal of Philosophy* 26: 267–82.

—— (1996*b*). 'The Role of Contrast in Causal and Explanatory Claims', *Synthese* 107: 395–419.

—— (2001). 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98: 273–99.

—— (2003). 'Of Humean Bondage', *British Journal for the Philosophy of Science* 54: 1–25.

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

HITCHCOCK, C., and KNOBE, J. (forthcoming). 'Cause and Norm'.

KAHNEMAN, D., and MILLER, D. (1986). 'Norm Theory: Comparing Reality to its Alternatives'. *Psychological Review* 80: 136–53.

—— and TVERSKY, A. (1982). 'The Simulation Heuristic', in Kahneman, Slovic, and Tversky (1982: 201–10).

—— SLOVIC, P., and TVERSKY, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

KIM, J. (1969). 'Events and Their Descriptions: Some Considerations', in Rescher (1969).

—— (1973). 'Causation, Nomic Subsumption, and the Concept of Event', *Journal of Philosophy* 70: 217–36.

KNOBE, J., and FRASER, B. (2008). 'Causal Judgement and Moral Judgment: Two Experiments', in Sinnott-Armstrong (2008: 441–8).

LEPORE, E., and MCLAUGHLIN, B. (eds.) (1985). *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell.

LEWIS, D. (1970). 'How to Define Theoretical Terms', *Journal of Philosophy* 67: 427–46.

—— (1973). 'Causation', *Journal of Philosophy* 70: 556–67.

—— (1983). 'New Work for a Theory of Universals', *Australasian Journal of Philosophy* 61: 343–77.

—— (1986*a*). *Philosophical Papers II*. Oxford: Oxford University Press.

—— (1986*b*). 'Postscripts to "Causation"', in Lewis (1986*a*: 172–213).

—— (1986*c*). 'Events', in Lewis (1986*a*: 241–69).

—— (1986*d*). 'Causal Explanation', in Lewis (1986*a*: 214–40).

—— (2000). 'Causation as Influence', *Journal of Philosophy* 97:182–97.

—— ([2000] 2004). 'Causation as Influence', in Collins, Hall, and Paul (2004: 75–106). Expanded version of Lewis (2000).

MACKIE, P. (1992). 'Causing, Delaying, and Hastening: Do Rains Cause Fires?' *Mind*

101: 483–500.

MASLEN, C. (2004). 'Causes, Contrasts, and the Nontransitivity of Causation', in Collins, Hall, and Paul (2004: 341–57).

MAUDLIN, T. (2004). 'Causation, Counterfactuals, and the Third Factor', in Collins, Hall, and Paul (2004: 419–44).

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

—— (2004). 'For Facts as Causes and Effects', in Collins, Hall, and Paul (2004: 309–24).

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

—— (2004). 'Difference-making in Context', in Collins, Hall, and Paul (2004: 139–80).

—— (2007). 'Causation in Context', in Price and Corry (2007: 191–223).

NORTON, J. (2007). 'Causation as Folk Science', in Price and Corry (2007: 11–44).

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

PRICE, H., and CORRY, R. (2007). *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Clarendon.

RESCHER, N. (ed.) (1969). *Essays in Honor of Carl G. Hempel*. Dordrecht: Reidel.

SANFORD, D. (1985). 'Causal Relata', in E. Lepore and B. McLaughlin (1985: 282–94).

SCHAFFER, J. (2006). 'Contrastive Causation', *Philosophical Review* 114: 297–328.

SINNOTT-ARMSTRONG, W. (ed.) (2008). *Moral Psychology*, ii. *The Cognitive Science of Morality*. Cambridge, Mass.: MIT.

SPIRTES, P., GLYMOUR, C., and SCHEINES, R. ([1993] 2000). *Causation, Prediction and Search*. New York: Springer.

STRAWSON, P. F. (1992). *Analysis and Metaphysics: An Introduction to Philosophy*. Oxford: Oxford University Press.

WOODWARD, J. (1984). 'A Theory of Singular Causal Explanation', *Erkenntnis* 21: 231–62.

—— (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

—— (2006). 'Sensitive and Insensitive Causation'. *Philosophical Review* 115: 1–50.

CHAPTER 18
CAUSES, LAWS,
AND ONTOLOGY

MICHAEL TOOLEY

DIFFERENT approaches to causation often diverge very significantly on ontological issues, in the case of both causal laws, and causal relations between states of affairs. In this chapter the main alternatives with regard to each will be set out.

1. CAUSATION: FUNDAMENTAL ONTOLOGICAL ISSUES

Causal concepts have surely been present from the time that language began, since the vast majority of action verbs involve the idea of causally affecting something. Thus, in the case of transitive verbs describing physical actions, there is the idea of causally affecting something external to one—one finds food, builds a shelter, sows seed, catches fish, and so on—while in the case of intransitive verbs describing physical actions, it is very plausible that they involve the idea of causally affecting one's own body—as one walks, runs, jumps, hunts, and so on.

It was not long after the very beginning of philosophy in ancient Greece that serious reflection concerning the nature of causation arose with Aristotle's famous discussion of causation in Book 2 of his *Physics*. The result was Aristotle's doctrine of four types (or, perhaps, aspects) of causes—material, formal, efficient, and final—an account that was immensely influential for about 2000 years.

What was not realized at any point during this time, however—perhaps because of the sense of familiarity with the idea of causation occasioned by the almost ubiquitous presence of causal concepts in even the most rudimentary parts of language—is that the concept of causation gives rise to very serious, puzzling, and difficult philosophical questions. Thus it was only many centuries after Aristotle, with David Hume and his famous discussions of the relation of cause and effect (1739–40 and 1748), that philosophers realized that the idea of causation was by no means simple and straightforward, and that there were problems both in offering an account of the meaning of that concept, and in specifying what the truthmakers could possibly be for causal propositions.

Why did Hume see what so many thoughtful philosophers before him had not? The reason, it would seem, was that Hume held—as did the other British empiricists, Locke and Berkeley —that while some concepts can be analysed in terms of other concepts, in the end analysis must terminate in ideas that apply to things in virtue of objects having properties and standing in relations that can be immediately given in experience. Hume therefore asked whether the relation of causation was one that was given in immediate experience. His conclusion was that

it was not. The question for Hume, accordingly, was how the concept of causation could be analysed in terms of ideas that do pick out properties and relations that are given in experience, and once this question was in view, Hume was able to show that arriving at a satisfactory answer was a very difficult matter.

One of the central issues in the philosophy of causation concerns, then, this Humean problem: is the concept of causation basic and unanalysable, or, on the contrary, does it stand in need of analysis? If it does need to be analysed, how can this be done?

When Hume tackled these questions, he focused upon causation as a relation between particular events, such as one billiard ball's hitting another. The analysis that he offered, however, of what it is for one particular event, or state of affairs, to cause another brought in the idea of a constant conjunction between two types of events under which the cause and the effect fell. What Hume referred to as a constant conjunction one might naturally refer to as a causal generalization, and a crucial question concerns what the relation is between causal relations between particular events and causal generalizations.

The idea of a causal generalization, however, covers quite different sorts of statements. On the one hand, there are statements that assert that events of one type are causally relevant to events of another type, without specifying what the causally efficacious properties are. 'Smoking causes lung cancer' is an example of such a statement, since there is no indication of what it is about smoking that is causally relevant to the development of lung cancer. Contrast this with a statement such as Newton's law of gravitation, which asserted that it was an object's having mass m_1, and being distance d from another object with mass m_2, that brought it about that each object was acted upon by a force equal to km_1m_2/d^2. Here one has a statement that not only asserts that there is *some* causal connection between states of affairs of certain types, but which also specifies what the causally relevant properties and relations are.

When causal generalizations function in the latter way, I shall speak of them as expressing causal laws. Moreover, since it seems plausible that causal generalizations that do not express causal laws obtain by virtue of underlying causal laws, it would seem that the basic ontological issues that arise in connection with causation can be set out by focusing upon causal generalizations that express causal laws.

If so, what are the basic ontological issues that need to be addressed if one is to arrive at a satisfactory account of the nature of causation? I would suggest that the following are central. First, how are causal laws and causal relations between states of affairs related? Are causal laws more basic, or causal relations between states of affairs, or are both ontologically basic? Secondly, how are causal facts related to non-causal facts? Are some causal states of affairs basic and irreducible, or are all causal facts logically supervenient upon, and reducible to, non-causal states of affairs? Thirdly, what sorts of properties and relations are involved in causal laws, and in causal relations?

2. CAUSAL LAWS AND CAUSAL RELATIONS: SINGULARIST, ANTI-SINGULARIST, AND INTERMEDIATE APPROACHES TO THE RELATION OF CAUSATION

Let us begin with the first of these questions. Among causal states of affairs, is it causal relations between events and states of affairs that are more basic, or is it causal laws? Or could it be that both causal relations and causal laws are basic?

If one considers familiar causal interactions, the idea that causal relations are more basic than causal laws has a strong initial appeal. A cue strikes a ball, and the ball rolls down the table. A glass is dropped, and shatters when it hits the floor. A dry cloth is placed in water, and emerges soaking wet. In these, as in a multitude of everyday occurrences, it is natural to say that one can just see the causing of one event by another, and that one can do this without knowing of any law under which these events fall. But, if this is so, must not causal relations be more basic than causal laws?

According to singularist approaches to causation, this is the correct view of the matter. Just as temporal priority is a relation between particular events, and spatial betweenness a relation between particular states of affairs, or objects at a time, so causation is, at bottom, just a relation between particular events.

It certainly appears to be true in the actual world, of course, that causal relations fall under general patterns, so that if a certain glass shatters when it hits the floor, similar glasses will also shatter when dropped from the same height onto floors composed of the same material. But, according to the singularist view, one particular event's causing another does not presuppose that that relation falls under any generalization. So, for example, as G. E. M. Anscombe ([1971] 1993: 92) puts it:

> If *A* comes from *B*, this does not imply that every *A*-like thing comes from some *B*-like thing or set-up or that every *B*-like thing or set-up has an *A*-like thing coming from it; or that given *B*, *A* had to come from it, or that given *A*, there had to be *B* for it to come from. Any of these may be true, but if any is, that will be an additional fact, not comprised in *A*'s coming from *B*.

Another vigorous advocate of a singularist approach to causation was C. J. Ducasse (1926: 61):

> The supposition of recurrence is thus wholly irrelevant to the meaning of cause; that supposition is relevant only to the meaning of law. And recurrence becomes related at all to causation only when a law is considered which happens to be a generalization of facts themselves individually causal to begin with. A general proposition concerning such facts is, indeed, a causal law, but it is not causal because general. It is general, i.e. a law, only because it is about a class of resembling facts; and it is causal only because each of them already happens to be a causal fact individually and in its own right (instead of, as Hume would have it, by right of its co-membership with others in a class of pairs of successive events).

This is a very natural view—though whether it is true is another question. In any case, we have here a crucial threefold division into those approaches to causation that view causal relations between states of affairs as primary, those that view causal laws as primary, and

those that adopt an intermediate view.

Approaches of the first sort accept the claim:

The Singularist Thesis with regard to Causal Relations. Events can be causally related without its being the case that those relations fall under any law.

In contrast with singularist approaches there are, at the other extreme, what I shall refer to as 'anti-singularist' approaches, according to which causal laws are more basic than causal relations. These involve acceptance of the claim:

The Anti-Singularist Thesis with regard to Causal Relations. Causal relations between events are logically supervenient upon, and reducible to, the totality of instances of non-causal properties and relations, together with causal laws. Accordingly, any two worlds that agree both with respect to all the non-causal properties of, and relations between, particulars, and with respect to all causal laws, must also agree with respect to all causal relations between states of affairs.

It is possible, however, to reject both these theses, and to accept, instead, this position:

The Intermediate View of Causal Relations. All causal relations between states of affairs are instances of laws, but such relations do not logically supervene upon the totality of instances of non-causal properties and relations, together with causal laws.

Since the time of Hume, most philosophers who have reflected upon the nature of causation have held that causal relations are reducible to causal laws plus the non-causal properties of, and relations between, events. But a number of philosophers have defended singularist views. Anscombe and Ducasse were mentioned above; two more recent defenders are Nancy Cartwright (1989) and Michael Tooley (1990).

Later, we shall look in detail at singularist views, and there we shall see that while they agree that causal relations are more basic than causal laws, they can differ radically concerning the ontology involved in causal relations.

3. Causal and Non-Causal States
of Affairs: Reductionism versus Non-Reductionism

Let us now turn to the second of the fundamental questions concerning the metaphysics of causation that was mentioned above, and which concerns the relation between causal facts and non-causal facts. The answers that philosophers have offered to this second question give rise

to another great divide concerning the ontology of causation—namely, that between reductionist and non-reductionist approaches, where this is a matter of the acceptance, or rejection, of the thesis:

> *Reductionism with respect to Causation*. Both causal relations between events, and causal laws, are logically reducible to states of affairs that involve only non-causal properties and relations. Accordingly, any two possible worlds that agree with respect to all the non-causal properties of, and relations between, entities of every sort must also agree with regard to all causal relations between states of affairs, and all causal laws.

3.1 Major Divisions within Reductionism

Acceptance of this general reductionist thesis is compatible, however, with very different views on the metaphysics of causation. First of all, there is the singularist/non-singularist divide. According to the reductionist who is a singularist, causal relations between events are reducible to the non-causal properties of, and relations between, particulars, whereas according to the reductionist who is not a singularist, the reduction of causal relations between events involves not only non-causal properties of, and relations between, particulars, but also causal laws.

Secondly, reductionists differ with regard to the accounts that they offer of the nature of laws. Most reductionists follow Hume in equating laws with cosmic regularities, or a restricted subset thereof. So, for example, one important possibility here is a view originally advanced by John Stuart Mill (1875: Bk. 2, ch. 4), rediscovered by Frank Ramsey in 1928 ([1928] 1990), and then popularized by David Lewis, who puts the view thus: 'We can restate Ramsey's 1928 theory of lawhood as follows: a contingent generalization is a *law of nature* if and only if it appears as a theorem (or axiom) in each of the true deductive systems that achieves a best combination of simplicity and strength' (1973 *b*: 73). Causal reductionists who do take a reductionist view of laws are committed to the thesis:

> *Strong Reductionism with respect to Causal Laws*. Any two worlds that agree with respect to all the non-causal properties of, and relations between, particulars, must also agree with respect to causal laws. Causal laws are, then, logically supervenient upon the totality of instances of non-causal properties and relations.

Reductionist accounts of laws are, however, open to serious objections. First, reductionist views typically entail that whether a certain law of nature exists depends upon how many instances the corresponding regularity has, and this is counterintuitive. Secondly, if one considers a probabilistic world containing two types of particles that happen never to interact, it does not seem to be logically impossible that there is a basic law concerning what would happen if they did interact. Reductionist views entail, however, that this is logically

impossible. Thirdly, it can be shown that no finite body of evidence, however great, can ever make it likely that an exceptionless generalization with an infinite number of instances is true, unless there is some atomic state of affairs that entails the generalization (Tooley 1987: 132–6). Fourthly, reductionist views entail, in the case of probabilistic laws, that worlds with different probabilistic laws—no matter how close the probabilistic laws are—could never have matching histories. Again, this seems deeply counterintuitive.

The controversy between reductionist and non-reductionist approaches to laws of nature is far from being settled, as is shown by current discussions (Carroll 2004). But in view of objections such as the above, some reductionists with regard to causation have embraced metaphysically more robust accounts of laws of nature, such as the view that laws involve irreducible second-order relations between universals (Dretske 1977; Tooley 1977; Armstrong 1983), or the view that laws involve ultimate dispositional properties of worlds (Ellis and Lierse 1994). The result is then this, weaker thesis:

> *Moderate Reductionism with respect to Causal Laws*. Any two worlds that agree both with respect to all the non-causal properties of, and relations between, particulars, and with respect to all laws of nature, must also agree with respect to causal laws. Causal laws are, then, logically supervenient upon the totality of instances of non-causal properties and relations, together with laws of nature.

Some philosophers—most notably Adrian Heathcote and David Armstrong (1991)—have suggested that all basic laws of nature are causal laws. If this view were correct, then moderate reductionism with respect to causal laws would be ruled out. But if one considers a Newtonian world, Newton's Third Law of Motion, which states that if body *A* exerts a force on body *B*, then body *B* exerts an equal and opposite force on body *A,* does not appear to be derivable from other Newtonian laws, and it is certainly not a causal law. So it appears that there can be basic, non-causal laws. Moreover, just as a Mill–Ramsey–Lewis account of the nature of laws is perfectly compatible with the possibility of underived laws of coexistence, so the same is true of at least some metaphysically more robust views of laws—such as the view that laws are second-order relations between universals.

Thirdly, and perhaps most important of all, reductionists differ with regard to the relevant non-causal properties and relations, and the resulting non-causal states of affairs that constitute the reductionist base upon which both causal laws and causal relations between states of affairs logically supervene. Especially crucial here is the divide between those who hold that the reductionist base must be restricted to Humean states of affairs, and those who hold that, on the contrary, non-Humean states of affairs are admissible. So let us turn now to that distinction.

What is a Humean state of affairs? The basic idea here is that Humean states of affairs are ones that satisfy Hume's thesis that there are no necessary connections between distinct existences. But how can such Humean states of affairs be defined? David Lewis, in characterizing the idea of Humean supervenience, offered this answer:

> Humean supervenience is named in honor of the greater denier of necessary connections. It is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact,

just one little thing and then another. (But it is no part of the thesis that these local matters are mental.) We have geometry: a system of external relations of spatiotemporal distance between points. Maybe points of spacetime itself, maybe point-sized bits of matter or aether or fields, maybe both. And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. For short: we have an arrangement of qualities. And that is all. There is no difference without difference in the arrangement of qualities. All else supervenes on that. (Lewis 1986 *b*: pp. ix–x)

According to this account, Humean states of affairs satisfy the conditions:

> (1) Humean states of affairs all consist of basic particulars having intrinsic properties and/or standing in relations.

> (2) All basic particulars are either point-sized objects, or points of spacetime.

> (3) The intrinsic properties in question are ones whose instantiation does not entail the existence of any basic particular other than the one in which the property is instantiated.

> (4) The only external relations between basic particulars are spatio-temporal distance relations.

Humean states of affairs, thus characterized, will certainly satisfy the constraint that distinct ones will not stand in any necessary relations to one another. But Lewis's account seems narrower than necessary. First of all, it does not seem necessary to restrict the relations to spatio-temporal distance relations. One could, for example, allow the relation of temporal priority—as Hume himself did. Or, more generally, one could allow any external relation. Secondly, the idea of Humean states of affairs seems perfectly compatible with there not being any point-sized entities or points of spacetime. A world consisting simply of infinitely divisible entities could still be a world of Humean states of affairs, as long as the intrinsic properties were all such that a given property's being instantiated did not entail the existence of anything that had a part that did not overlap with the entity in which the property was instantiated.

I am inclined to think, then, that Humean states of affairs can be characterized more broadly, and more simply, thus:

> (1) Humean states of affairs all consist of particulars having intrinsic properties and/or standing in external relations.

> (2) All intrinsic properties of particulars are ones whose instantiation does not entail the existence of any particular that has a part that does not overlap the particular in which the property is instantiated.

What would be an example of a non-Humean state of affairs? Consider a state of affairs that

consists of the existence, at some time, of an indestructible object. That state of affairs entails the existence of that object at all later times. Does that mean that one has here a non-Humean state of affairs? The answer depends upon what account one offers of what it is to be indestructible. Suppose, on the one hand, that one holds that dispositional 'properties' are logically supervenient upon categorical properties plus laws of nature, so that an object is indestructible because it has some categorical, intrinsic property, *P,* such that it is a law of nature that anything that has property *P* at any time will exist at all later times. Then no non-Humean state of affairs need be involved, since the possession of the categorical property *P* does not logically entail the existence of the entity in question at later times. But suppose, on the other hand, that one holds instead that dispositional properties are basic and irreducible, so that an object is indestructible because it possesses a basic property *D* such that anything that has property *D* at a time will necessarily exist at all later times. Then something's having property *D* is a non-Humean state of affairs, since the existence of an object with the intrinsic property *D* at one time logically entails the existence of that object at all later times.

3.2 Major Divisions within Non-Reductionism

The choices confronting non-reductionists parallel very closely, with two exceptions, those that arise for reductionists. First, non-reductionists with regard to causation can adopt either a singularist or a non-singularist view. Secondly, non-reductionists can identify laws with (certain sorts of) cosmic regularities, or they can adopt a metaphysically robust view of laws. Thirdly, they can hold either that causation involves non-Humean states of affairs, or that, while one may have to go beyond the states of affairs that Hume himself judged admissible, the extra states of affairs that are needed are all Humean in the sense defined above, since they are all local states of affairs that do not introduce any necessary connections between distinct existences.

One point of divergence as regards the main options open to reductionists and non-reductionists is that in the case of *Humean,* non-reductionist approaches to causation, there would seem to be only a singularist option. The reason is that a non-reductionist, non-singularist account of causation requires a metaphysically robust conception of laws of nature, according to which non-probabilistic laws logically entail, but are distinct from, regularities, and laws, thus conceived, are non-Humean states of affairs.

The second point of divergence as regards the main options open to reductionists and non-reductionists is that in the case of *non-Humean,* non-reductionist approaches to causation, there are two distinct types of non-singularist alternatives open to one. First, there is the anti-singularist view, according to which causal relations are reducible to non-causal properties of, and relations between, particulars, together with causal laws, and, secondly, there is the intermediate view, according to which causal relations are not thus reducible, even though for every causal relation between events there must be a corresponding causal law under which that relation falls as an instance.

3.3 Alternative Views: The Main Options

Summing up, then, the main, high-level options with regard to the ontology of causation are as follows. First, there are Humean reductionist approaches, which can take both singularist and anti-singularist forms. Secondly, there are non-Humean reductionist approaches, which can also take both singularist and anti-singularist forms. Thirdly, there is a singularist, Humean non-reductionist approach. Finally, there are non-Humean non-reductionist approaches, which can take singularist, anti-singularist, and intermediate forms.

4. HUMEAN REDUCTIONIST APPROACHES TO CAUSATION: SOME IMPORTANT ALTERNATIVES

Let us now consider briefly these possibilities, beginning with Humean reductionist approaches. What are some of the main ways, then, in which one might attempt to formulate a reductionist account of causation that does not go beyond an ontology consisting of Humean states of affairs? In answering this question, let us begin by focusing first on anti-singularist approaches, according to which causal relations logically supervene upon, and are reducible to, laws together with the non-causal properties of, and relations between, particulars.

4.1 Anti-Singularist, Humean Reductionist Approaches

Such approaches to causation go back to one of David Hume's own definitions of a cause —'an object precedent and contiguous to another, and where all the objects resembling the former are placed in a like relation of priority and contiguity to those objects, that resemble the latter' (1739–40: 1. 3. 14)—since it seems reasonable to interpret Hume's requirement of a constant conjunction of relevant types of objects as expressing a law-based connection.

Following Hume, a number of later philosophers advanced variants on his approach that made central use of the idea of nomological conditions—either necessary, or sufficient, or both—in an attempt to analyse the relation of causation. (For an account of some approaches of this sort, see sects. 2 and 3 of the Introduction to Sosa and Tooley 1993: 5–9.)

But, especially with the advent of quantum mechanics, it has become very plausible that no account of causation can be satisfactory unless it is compatible with the idea of probabilistic causal laws. As a result, among the most important present-day, anti-singularist, Humean, reductionist approaches are ones that replace Hume's constant conjunction requirement with the requirement that a cause must stand in the relation of positive statistical relevance to its effect, where this is cashed out in terms of a relation between conditional and unconditional relative frequencies. Approaches of this sort have been advanced and defended, for example, by Hans Reichenbach (1956), I. J. Good (1961 and 1962), Patrick Suppes (1970), and many other more recent philosophers.

Another very different, and very influential, anti-singularist, Humean, reductionist account, however, is the counterfactual approach to causation set out and defended by David Lewis ([1973*a*] 1986; 1986*a*). His basic strategy involves analysing causation using a narrower

notion of causal dependence and then analysing causal dependence counterfactually: (1) an event *c* causes an event *e* if, and only if, there is a chain of causally dependent events linking *e* with *c*; (2) an event *g* is causally dependent upon an event *f* if, and only if, had *f* not occurred, *g* would not have occurred. See Ch. 8 above.

Causes, so construed, need not be necessary for their effects because counterfactual dependence, and hence causal dependence, is not necessarily transitive. Nevertheless, Lewis's approach is closely related to necessary-condition analyses of causation since the more basic relation of causal dependence is a matter of one event's being counterfactually necessary in the circumstances for another event.

Some well-known approaches to counterfactuals offer analyses that involve causal notions (Jackson 1977; Kvart 1986). If there were no alternative to such analyses, any counterfactual analysis of causation would either be circular, or an analysis of the relation of causation in terms of causal laws, and an account would then need to be given of the latter. But Lewis held that the correct analysis of counterfactuals was a Stalnaker/Lewis analysis in terms of similarity relations across possible worlds (Stalnaker 1968; Lewis 1973*b*; [1979] 1986. For arguments for the view that this sort of approach is untenable, see, for example, Kvart 1986 and Tooley 2003).

Why is Lewis's account of causation an anti-singularist account, given that it contains no explicit reference to laws of nature? The reason is that counterfactual dependence is to be analysed in terms of the similarity of possible worlds: 'If *c* had not occurred, *e* would not have occurred' is true if and only if *e* fails to occur at the 'closest' possible world (that is to say, the possible world most similar to the actual world) at which *c* fails to occur. If there were no law relating *C*-type events and *E*-type events, either directly or indirectly, there would be no reason to expect the counterfactual that grounds the causal relation between *c* and *e* to come out true, since some possible world that contained event *e*, but lacked event *c*, would surely be closer to the actual world—which does contain events *c* and *e*—than any world where both *c* and *e* were absent.

4.2 A Singularist, Humean Reductionist Approach

Let us now turn to Humean reductionist approaches of a singularist sort, according to which events can be causally related without its being the case that those relations fall under any law. Very few philosophers indeed have defended this type of account, but C. J. Ducasse was one who did. Ducasse was impressed by Hume's complaint about his own first (regularity) definition of causation—to the effect that it was 'drawn from circumstances foreign to the cause' and 'from something extraneous to it' (1748: 7 2). Ducasse (1926: 59), accordingly, eliminated Hume's reference to constant conjunction, thereby generating an account of causation that does not look beyond the actual situation involving the cause and its effect:

Considering two changes, *C* and *K* (which may be either of the same or of different objects), the change *C* is said to have been sufficient to, i.e. to have caused, the change *K* if:

1. The change *C* occurred during a time and through a space terminating at the instant *I* at

the surface S.

2. The change K occurred during a time and through a space beginning at the instant I at the surface S.
3. No change other than C occurred during the time and through the space of C, and no change other than K during the time and through the space of K.

(See Ch. 7 for further discussion of Ducasse's view.)

5. NON-HUMEAN REDUCTIONIST APPROACHES TO CAUSATION: TWO ALTERNATIVES

5.1 A Singularist Approach: Ultimate Dispositional Properties

The options are much more limited when one turns from Humean reductionist approaches to non-Humean ones. On the singularist side, the only alternative—but a very important one—involves attempting to reduce causal relations between events to the combination of non-causal properties and relations together with dispositional properties.

Initially, such an approach may seem to have things backwards, since there is a very simple and natural analysis of dispositional concepts in terms of causal laws. So, for example, this seems like a very plausible analysis of the concept of being water-soluble:

A is water-soluble at time t^* = df. There is some causal law L and some intrinsic property P such that A has property P at time t^* and L entails that for any X and any time t, if X has property P at time t and X is in water at time t, then that state of affairs causes it to be the case that X is dissolving throughout some temporal interval immediately following t.

Philosophers who maintain that causal relations are logically reducible to non-causal properties and relations together with dispositional properties hold, however, that the above sort of analysis of dispositional concepts is unsound, and that dispositional concepts instead pick out basic, intrinsic properties of objects.

Thus a number of philosophers—including Rom Harré and Edward Madden (1975), Nancy Cartwright (1989), C. B. Martin (1993), Ellis (2001), and Molnar (2003)—have both advocated an ontology according to which there are *irreducible* dispositional properties, powers, propensities, chances, and the like, and maintained that such an ontology enables one to provide an account of causation. Often, however, the details have been rather sparse, perhaps because if one focuses upon non-probabilistic cases, the correct analysis may seem quite straightforward. Consider, for example, the statement that A's being in water caused it to dissolve. A very simple and natural analysis of this is

A's being in water at time t^* caused it to dissolve throughout a temporal interval

immediately after t^* = df. There is some dispositional property (or set of dispositional properties), D, such that A had property D at time t^*, and for any X and any time t, X's being in water at time t and X's having property D at time t logically entails that X is dissolving throughout a temporal interval immediately after t.

But when one considers probabilistic dispositions, or propensities, one no longer has a relation of logical entailment, and so a more complex account is needed. Clear analyses of causation in terms of objective chances were, however, set out in 1986 both by D. H. Mellor (1986) and by David Lewis (1986*a*) and then, more recently, Mellor has offered a very detailed statement and defence of this general approach in his book *The Facts of Causation* (1995).

Mellor's approach, in brief, is roughly as follows. First, Mellor embraces an ontology involving objective chances, where the latter are ultimate properties of states of affairs, rather than being logically reducible to causal laws together with non-dispositional properties, plus relations. Secondly, Mellor proposes that chances can be defined as properties that satisfy three conditions: (1) the Necessity Condition: if the chance of P's obtaining is equal to 1, then P is the case; (2) the Evidence Condition: if one's total evidence concerning P is that the chance of P is equal to k, then one's subjective probability that P is the case should be equal to k; (3) the Frequency Condition: the chance that P is the case is related to the corresponding relative frequency in the limit. Thirdly, chances enter into basic laws of nature. Fourthly, Mellor holds that even basic laws of nature need not have instances, thereby rejecting reductionist accounts in favour of a non-reductionist view. Fifthly, any chance that P is the case must be a property of a state of affairs that temporally precedes the time at which P exists, or would exist. Finally, and as a very rough approximation, a state of affairs C causes a state of affairs E if and only if there are numbers x and y such that (1) the total state of affairs that exists at the time of C—including laws of nature—entails that the chance of E is x, (2) the total state of affairs that would exist at the time of C, if C did not exist, entails that the chance of E is y, and (3) x is greater than y.

Why are analyses of causation in terms of ultimate, irreducible dispositional properties cases of non-Humean reductionist accounts? The answer is that, given the existence of such ultimate dispositional properties, there can be logical connections between distinct states of affairs—such as, on the one hand, the conjunctive state of affairs that consists of A's being in water at time t and A's having the ultimate dispositional property of water-solubility at time t, and, on the other hand, the temporally distinct state of affairs that consists of A's dissolving throughout some temporal interval just after time t.

5.2 An Anti-Singularist, Non-Humean, and Reductionist Approach: Strong Laws

Let us now consider the idea of an anti-singularist, non-Humean, reductionist account of causation. Here, too, there seems to be only one general sort of possibility—namely, that represented by a view proposed by David Armstrong and Adrian Heathcote (Heathcote and

Armstrong 1991; Armstrong 1997: ch. 14).

Their view involves a number of elements, two of which are as follows. First, when one has two states of affairs that are causally related, there is nothing extrinsic to the causal process in question that is needed for the causal sequence to obtain. Secondly, and as Anscombe ([1971] 1993) contended, there does not seem to be any purely conceptual or a priori argument that shows that causal relations must be instances of laws.

These two points suggest a singularist view of causation. But Armstrong and Heathcote proceed to develop an anti-singularist account. First of all, they contend that no regularity view of laws of nature is satisfactory, and that, in particular, one needs to embrace metaphysically robust laws involving an irreducible second-order relation of nomic necessitation between universals. Secondly, whenever there is an instance of a law, that second-order relation that holds between the relevant universals must be present in the instances of the law. Thirdly, they suggest that the reason anomic causation is not to be found in our world is that the relation of causation is identical with the relation of nomic necessitation. Finally, they suggest that this identity is necessary a posteriori.

The upshot is that while causal relations between states of affairs do not logically supervene upon, and are not conceptually reducible to, instances of strong laws of nature, they metaphysically supervene upon such instances, and are ontologically reducible to them. So while anomic causation is conceptually possible, it is metaphysically impossible.

6. NON-REDUCTIONIST APPROACHES TO CAUSATION

Finally, let us now turn to non-reductionist approaches to causation.

6.1 Singularist Accounts and the Direct Observability Claim

In the case of singularist accounts, there are two very different possibilities. The first has been defended by Elizabeth Anscombe ([1971] 1993) and Evan Fales (1990), and the basic contentions involved here are, first, that causal relations between states of affairs are, in favourable conditions, directly or immediately observable; secondly, that the concept of causation is conceptually basic; and, thirdly, that causation can therefore be classified as a basic, irreducible relation.

Anscombe, in her discussion, appeals to a variety of everyday situations involving the perception of causal relations, such as seeing a knife cut through butter. Fales, realizing that it can be argued that inference is involved in such cases, and thus that causation is not in such cases *immediately* observable, focuses instead upon introspective awareness of one's own acts of willing, and—following Armstrong (1968)—tactile awareness of pressure upon one's body. See Ch. 22 below.

If one can, in a single act of perception, directly observe that two states of affairs are

causally related, then, unless the a posteriori identity proposed by Armstrong and Heathcote obtains, there would seem to be good reason for holding that a singularist approach to causation is correct, since, on the one hand, the idea that a single act of perception could justify the conclusion that a certain occurrence fell under some law or other does not seem plausible, and, on the other, precisely this must be the case if the existence of a causal relation between two states of affairs entails the existence of some corresponding law.

Similarly, and unless the Armstrong/Heathcote view is correct, if the concept of causation is conceptually primitive and unanalysable, the occurrence of causally related states of affairs cannot require the existence of any corresponding law of nature.

Finally, this approach, while it involves an ontological claim that Hume argued against, does not involve any non-Humean states of affairs in the sense defined above, since the instantiation of an irreducible, singularist relation of causation would not entail any logical connections between distinct existences.

6.2 An Anti-Singularist, Non-Reductionist Approach to Causation

Before considering a second singularist, non-reductionist approach to causation, it will be helpful to consider a certain closely related anti-singularist approach. Three ideas are central to the anti-singularist approach in question. First, laws are to be identified with states of affairs consisting of irreducible, second-order relations between universals. Secondly, an account of the nature of laws in general, contrary to the view of Armstrong and Heathcote, does not on its own provide one with an account of the nature of *causal* laws. Thirdly, the theoretical definition of causal laws involves certain probability relations.

What reasons are there for thinking that causal laws cannot be identified with *basic* laws, as on the Armstrong/Heathcote view? Consideration of a Newtonian world provides at least two reasons. First, and as was mentioned earlier, Newton's Third Law of Motion is not a causal law, nor is it derivable from the other laws of Newtonian physics. It is, then, both basic and non-causal. Secondly, Newton's laws are time-symmetric, so that, if this were a Newtonian world, the complete state of the universe in the year 1900 would nomically necessitate the state of the universe in the year 2000, but it would also nomically necessitate the state of the universe in the year 1800. The latter, however, is not a causal relation, so nomic necessitation cannot be equated with causal necessitation. More generally, and Newtonian worlds aside, given any conservation law, the relevant state of the world—such as how much mass or charge or spin it contains—nomically necessitates not only later states but also earlier ones.

How, then, are causal laws to be distinguished from non-causal laws? The basic idea (Tooley 1987) is that causal laws can be defined as states of affairs involving second-order relations between universals that satisfy certain probability postulates—two of the most crucial of which may be stated, in slightly simplified form, as follows, where $C(P, Q)$ is an abbreviation for 'It is a causal law that for all x, and all t, if x has property P at time t, then x has property Q throughout some temporal interval immediately after t':

(P_1) : $\text{Prob}(Px/C(P, Q)) = \text{Prob}(Px)$

$$(P_2) : \text{Prob}(Qx/C(P, Q)) = \text{Prob}(Px) + [\text{Prob}(\sim Px)\times$$
$$\text{Prob}(Qx/\sim Px)]$$

What these two postulates say is that, given the information that states of affairs of type *P* cause states of affairs of type *Q*, but given no additional information, the posterior probability of a state of affairs of type *P* is precisely equal to the prior probability of states of affairs of type *P*, whereas, by contrast, the posterior probability of a state of affairs of type *Q*, given only the information that states of affairs of type *P* cause states of affairs of type *Q*, is a function of the prior probability of states of affairs of type *P*. The posterior probability of an effect, accordingly, is a function of the prior probability of its cause, whereas the posterior probability of a cause is not a function of the prior probability of its effect.

Causal laws can then be defined as states of affairs consisting of irreducible, second-order relations between universals that satisfy the above postulates, and this non-reductionist account of causal laws can then be combined with an account of causal relations between states of affairs according to which they logically supervene upon, and are reducible to, causal laws plus non-causal properties of, and relations between, particulars. The result is an anti-singularist, non-reductionist account of causation.

6.3 An Alternative Singularist Account: Causation as an Irreducible, Theoretically Specified Relation

The problem is that there are strong arguments that show that no anti-singularist analysis of causation can possibly be sound (Foster 1979; Armstrong 1983; Tooley 1987), since one can show that there are possible worlds that agree with respect to all non-causal properties of, and relations between, particulars, all dispositions and propensities, all causal and non-causal laws, and the direction of time, but that disagree with respect to some causal relations between states of affairs. So neither any anti-singularist account, nor any singularist account that reduces causal relations to non-causal facts, including ones involving dispositions, can be sound.

One possibility at this point is to move to an intermediate account, according to which causal relations presuppose corresponding causal laws, even though they do not supervene upon causal laws plus non-causal properties and relations (Tooley 1987). Such a view is, however, hard to motivate. Moreover, it turns out that it is possible to modify the anti-singularist account just given to get a singularist account (Tooley 1990), and that seems to be a much more promising route.

The account in question can be arrived at by reinterpreting the expression '*C(P, Q)*' in postulates (*P*1) and (*P*2) so that the proposition that '*C(P, Q)*' expresses is that it is a law that for all *x*, if *x* has property *P*, then *x*'s having property *P causes x* to have property *Q*. For when causal laws are formulated in a way that involves explicit reference to the relation of causation, one can then go on to define causation as the unique relation between states of affairs that is such that any laws into which that relation enters in a certain way must satisfy

the postulates in question. One then has a singularist account, for although the concept of causation has been analysed in a way that involves the idea of laws of nature, the analysis does not entail that causal relations between states of affairs need fall under any laws: anomic causation is logically possible.

FURTHER READING

For a collection of articles covering most of the basic approaches to causation, see Sosa and Tooley (1993). To explore the issue of reductionism versus non-reductionism as it arises in connection with the closely related question of the nature of laws, see Carroll (2004). Finally, Dowe and Noordhof (2004) contains many excellent articles on both probabilistic and counterfactual approaches to causation.

REFERENCES

ANSCOMBE, G. E. M. ([1971] 1993). *Causality and Determination*. Cambridge: Cambridge University Press; repr. in Sosa and Tooley (1993: 88–104).

ARMSTRONG, D. M.(1968). *A Materialist Theory of Mind*. London: Routledge & Kegan Paul.

—— (1983). *What Is a Law of Nature?* Cambridge: Cambridge University Press.

—— (1997). *A World of States of Affairs*. Cambridge: Cambridge University Press.

CARROLL, J. W. (ed.) (2004). *Readings on Laws of Nature*. Pittsburgh: University of Pittsburgh Press.

CARTWRIGHT, NANCY (1989). *Nature's Capacities and their Measurement*. Oxford: Clarendon.

DOWE, P., and NOORDHOF, P. (eds.) (2004). *Cause and Chance: Causation in an Indeterministic World*. London: Routledge.

DRETSKE, FRED I. (1977). 'Laws of Nature', *Philosophy of Science* 44/2: 248–68.

DUCASSE, C. J. (1926). 'The Nature and the Observability of the Causal Relation', *Journal of Philosophy* 23: 57–68; repr. in Sosa and Tooley 1993: 125–36.

ELLIS, B. (2001). *Scientific Essentialism*. Cambridge: Cambridge University Press.

—— and LIERSE, C. (1994). 'Dispositional Essentialism', *Australasian Journal of Philosophy* 72, 27–45.

FALES, E. (1990). *Causation and Universals*. London: Routledge.

FOSTER, J. (1979). 'In *Self*-Defence', in G. F. Macdonald (ed.), *Perception and Identity*. London: Macmillan.

GOOD. I. J. (1961 and 1962). 'A Causal Calculus', Parts 1 and 2, *British Journal for the Philosophy of Science* 11 (1961): 305–18; 12 (1962): 43–51.

HARRÉ, R., and MADDEN, E. M. (1975). *Causal Powers: A Theory of Natural Necessity*. Oxford: Blackwell.

HEATHCOTE, A., and ARMSTRONG, D. M. (1991). 'Causes and Laws', *Noûs* 25: 63–73.

HUME, D. (1739–40). *A Treatise of Human Nature*. London.

—— (1748). *An Enquiry Concerning Human Understanding*. London.

JACKSON, F. (1977). 'A Causal Theory of Counterfactuals', *Australasian Journal of*

Philosophy 55: 3–21.

KVART, I. (1986). *A Theory of Counterfactuals*. Atascadero, Calif.: Ridgeview.

LEWIS, D. K. ([1973*a*] 1986). 'Causation', *Journal of Philosophy* 70: 556–67; repr. in Lewis (1986*b*: 159–72).

—— (1973*b*). *Counterfactuals*. Cambridge, Mass.: Harvard University Press.

—— ([1979] 1986). 'Counterfactual Dependence and Time's Arrow', *Noûs* 13: 455–76; repr. with postscripts, in Lewis (1986*b*: 32–66).

—— (1986*a*). 'Postscripts to "Causation"', in Lewis (1986*b*: 172–213).

—— (1986*b*). *Philosophical Papers II*. Oxford: Oxford University Press.

MARTIN, C. B. (1993). 'Power for Realists', in J. Bacon, K. Campbell, and L. Reinhardt (eds.), *Ontology, Causality and Mind*. Cambridge: Cambridge University Press.

MELLOR, D. H. (1986). 'Fixed Past, Unfixed Future', in B. Taylor (ed.), *Contributions to Philosophy: Michael Dummett*. The Hague: Nijhoff, 166–86.

—— (1995). *The Facts of Causation*. London: Routledge.

MILL, J. S. (1875). *A System of Logic*. London: Longmans.

MOLNAR, G. (2003). *Powers: A Study in Metaphysics,* ed. S. Mumford. Oxford: Oxford University Press.

RAMSEY, F. P. ([1928] 1990). 'Universals of Law and of Fact', in D. H. Mellor (ed.), *F. P. Ramsey: Philosophical Papers*. Cambridge: Cambridge University Press.

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

SOSA, E., and TOOLEY, M. (eds.) (1993). *Causation*. Oxford: Oxford University Press.

STALNAKER, R. C. (1968). 'A Theory of Conditionals', in N. Rescher (ed.), *Studies in Logical Theory*. Oxford: Blackwell, 98–112.

SUPPES, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

TOOLEY, M. (1977). 'The Nature of Laws'. *Canadian Journal of Philosophy* 7: 667–98.

—— (1987). *Causation: A Realist Approach*. Oxford: Oxford University Press.

—— (1990). 'The Nature of Causation: A Singularist Account', in, D. Copp (ed.), *Canadian Philosophers, Canadian Journal of Philosophy* suppl. 16: 271–322.

—— (2003). 'The Stalnaker-Lewis Approach to Counterfactuals', *Journal of Philosophy* 100: 321–7.

CHAPTER 19
CAUSAL RELATA

DOUGLAS EHRING

WHAT kinds of things can be causes and effects, the relata of the causal relation? The most popular candidates are events and facts, but there are alternatives including tropes, exemplifications of universals, and objects. A range of causal phenomena, some of which are controversial, such as causation by and of absences, has been offered to decide this issue. Specifying the main candidates and considering arguments based on these phenomena will be my major, but not sole focus. There are also the questions of whether or not there is one kind of causal relatum to which all other types reduce and of how many relata the causal relation has (the received view being two, but three and four have also been proposed recently). I will also consider the relation between causal relata and causally relevant properties.

One way to make progress in determining the nature of the relata of a relation is by reference to various relevant aspects of that relation. I will follow that method here by working my way to an account of causal relata by reference to certain formal and substantial features of the causal relation: the (perhaps limited) transitivity and law-relative intrinsicness of the causal relation, spatial and temporal aspects of that relation, and, finally, the role of qualitative persistence in causation. As I sort through the main theories of causal relata, I will make use of these aspects of the causal relation. I will argue that the intrinsicness of the causal relation undermines the main case for facts as the causal relata, which is based on causation by and of absences. Furthermore, I will argue that since causes and effects are generally temporally and spatially related to each other, facts could not be causes and effects. I will also argue that the transitivity of causation rules out at least one major candidate for causal relata, coarse-grained events. And, finally, I will argue that since the best theory of causation employs the notion of qualitative or property persistence, the best candidate for causal relata must be based around tropes or particularized properties.

1. THE CANDIDATES

The nature of causal relata and of causes/effects is a single issue for those who hold the widely shared assumption that causation is always and everywhere a relation. However, for those philosophers who hold that causation is only sometimes a relation (e.g. Lewis) or never a relation (Mellor 1995: 165), causes/effects are either sometimes not causal relata (for Lewis, the causal relata 'go missing' in the case of causation by/of absences) or causes/effects are never causal relata (ibid.). In the following, however, I will move freely between 'causal relata' and 'causes/effects', except when it seems important to distinguish these issues. I will

begin by briefly cataloguing the main candidates for causal relata and, after that, discuss the issue of whether there is only one kind of causal relatum. I will then discuss the advantages and disadvantages of these various candidates. In the case of one candidate, objects, I will both describe and evaluate in the same section. My focus will be on the relata of token-level or singular causation—for example, Jones's smoking caused his cancer—rather than type-level causation—for example, smoking causes cancer. (For two general discussions of theories of causal relata, see Moore 2005 and Schaffer 2003*a*.)

1.1 Event Views

The most popular view of causal relata is that they are non-repeatable events including what we normally think of as events in the form of changes, but also including unchanging states. Within the event camp, theories vary primarily depending on how they individuate events—what conditions are specified under which an event e is or is not identical to an event e', with some theories making events more or less coarse-grained and others making events fine-grained to varying degrees. To illustrate the difference, a fine-grained theory might distinguish between the bolt's failure and the bolt's sudden failure whereas a coarse-grained theory might not.

The Coarse-Grained Event View. Causal relata are events and events are individual, concrete occurrences, spatio-temporally located particulars, with an indefinite number of properties. There are two main proposals for individuating events in a coarse-grained manner, causally and spatio-temporally. Under the first proposal, events e and e' are identical just in case they have the same causes and effects, leaving open the possibility of non-identical but exactly coinciding events (Davidson [1969] 1980: 179). A ball's turning and its heating up are different events since they have different causes and effects, even if they coincide spatio-temporally. Under the second proposal, events are individuated by their spatio-temporal locations, ruling out such coinciding events. (Quine 1985; Davidson 1985: 175). Given either way of individuating events, the same event can be specified by more than one description, and there is no built-in restriction as to which of an event's properties may be causally relevant to its causal relations.

The Kimian Fine-Grained Event View. Causal relata are events and events are exemplifications of universals (including *n*-adic universals) by a concrete object (or *n*-tuples of concrete objects) at a time (Kim [1973] 1993; [1976] 1993). In the monadic case, events c and c' are identical just in case their constitutive properties/objects/times are identical. The event of the bolt's failure with the constitutive property of *failure* is not identical to the event of bolt's sudden failure with the constitutive property of *sudden failure*. In contrast to the Quinean/later Davidsonian view, the same spatio-temporal region may include a large number of events and there *is* a restriction as to which of an event's properties may be causally relevant to its causal relations: an event's causal relations are determined by the property the exemplification of which by its constitutive object is the event, and not by properties merely exemplified by the event.

The Lewisian Event View. All causal relata are events, but events have essences—conditions that must be satisfied for that event to occur—and events that differ in their essential or accidental properties are not identical (Lewis 1986*a*: 245). John saying 'hello' loudly might involve at least two coinciding events. One is essentially a saying 'hello' loudly and the other is a saying 'hello'—loudly, but not essentially so. These two events occupy the very same region in the actual world, share all their categorical properties in the actual world, but they are not identical since they differ in their essential/accidental properties. A large number of events overlapping perfectly in a single spatio-temporal region of a world is a possibility, in contrast to the Quinean/later Davidsonian view. An event is modally fragile to the degree that it could not have occurred otherwise than how it did occur. If each of an event's properties is such that that very event could not have occurred without it, the event is extremely fragile.

1.2 Property Instance Views

On this view, causal relata are property instances, but what a property instance is will vary depending on what properties are. If properties are universals, then property instances are exemplifications of universals at a time by a particular. If properties are tropes, causal relata are tropes. Other possibilities depend on other theories of properties, and some proponents of this view leave it open what a property instance is (Paul [2000] 2004: 213). There is a further issue of whether or not causal relata are limited to a restricted class of property instances, no matter if property instances are exemplifications of universals or tropes, such that some property instances cannot be causal relata. For example, one might restrict the property instances that can be causal relata to property instances of objects (Honderich 1988: 15), or of events (aspects of events (Dretske 1977)), or make no such restrictions (Paul [2000] 2004; Ehring 1997). (Sanford (1985) holds that both events and event aspects are causal relata.)

Exemplication of Universals View. If properties are universals—capable of simultaneous, multiple location—then a property instance consists in the exemplification/instantiation of a universal by a particular at a time. Causal relata are such exemplifications of universals (Armstrong 1997: 205). How fine-grained causal relata are will depend on how universals are individuated, say as finely as necessarily coextensive predicates or sparsely as the properties of 'total science' (Schaffer 2003*a*: 8). (But note that the exemplification-of-a-universal view does not allow for causal relata consisting in the *non-exemplification* of a property by an object at a time as does Menzies (1989: 67); see also Barwise and Perry (1983) on real situations as causal relata.) This view collapses into the Kimian event view if the particulars that do the exemplifying include only objects and the only permissible universals are event-generating, and into the Dretske's aspect-of-an-event view if the particulars that do the exemplifying are restricted to events.

The Trope View. Causal relata are tropes or particularized properties, not exemplifications of universals and not universals since tropes are not capable of simultaneous, multiple location (Williams 1953; Campbell 1990; Ehring 1997). The cause of my current visual experience is a particular whiteness trope, not the exemplification of the universal *whiteness* by this page nor that universal itself. How fine-grained causal relata are will depend on how tropes are individuated. The trope view does not necessarily exclude events as causal relata if

events are reducible to tropes (for example, if events are sequences of tropes (Campbell 1990: 22). If the tropes that are causal relata are restricted to tropes that characterize events, we get a trope-version of the Dretske event aspect view.

1.3 The Fact View

Another alternative to the event view is that causes/effects are facts (Mackie 1974; Bennett 1988; 1995; Mellor 1995; 2004). (Events are causal relata, if they are at all, only derivatively, depending on the causal relations between facts about events.) What it will mean for facts to be causal relata depends on what facts are—true propositions or otherwise—and how they are individuated. As to the issue of individuation, facts might be individuated by the conditions under which two sentences are thought to express the same fact. For example, one might hold that two sentences express the same fact if and only if they are a priori interderivable. (Bennett (1988: 41) outlines other possibilities along these lines.) On the other hand, one might individuate facts causally: two sentences express the same fact if and only if their intersubstitution in any causal context would never change the truth value of that causal sentence (Mellor 2004: 313). There are two broad interpretations of what facts are.

The Propositional View. Facts are true propositions (Bennett 1988: 7) and they are individuated however propositions are individuated. Furthermore, facts understood as true propositions do not have, exist at, or obtain at specific times or locations. For example, the fact that New York is on the east coast of the USA does not exist on or have a location on the east coast. Neither does it obtain there rather than somewhere else. True propositions are about the world and are not occurrent features of the world with locations (Menzies 1989: 61).

The Non-Propositional View. Facts are not true propositions, but are whatever extra-propositional entities make true propositions true (Meinertsen 2000: 174). Mellor uses the term 'facta' for 'entities in our world, whatever they may be, whose existence or non-existence makes true statements true' (Mellor 1995: 162). Facts are the situations or states of affairs or exemplifications of properties or tropes or whatever that make true propositions true. In this sense, facts are non-propositional and have spatio-temporal locations. In the following, by 'fact' I will mean 'true proposition'. 'Facts' in at least some non-propositional varieties are covered indirectly by my treatment of the various non-propositional candidates for causal relata.

1.4 Objects and Persons

Some philosophers hold that at least some objects or substances can be causes, although not effects, and that object causation is not reducible to causation involving only non-object relata such as events or facts (Reid 1969; Chisholm 1966; Taylor 1966); Byerly (1979) allows both animate and inanimate object-causes, and Lowe (2002: 208–11) and Swinburne (2000) hold that *only* objects can be causes. Reid (1969) rules out non-person substance-causes). On this view, if object O causes event e that is not a matter of the event of O's having some property P causing e, for example. The main objection to objects-as-causes concerns the

timing of the effect. If it is John, and not John's having some property, that deterministically caused his ears to move today, then his ears should have moved yesterday, if he existed then too (Aune 1977: 6). A substance—rather than that substance's having some property—that exists before and after an effect begins cannot account for why that event begins at that time, but a cause can and does (or, in the indeterministic case, a cause explains why that event begins at that time to have a certain objective chance). There is some difference in properties in the object over time that is causally responsible for the time (or objective chance thereof) of the effect (Broad 1952: 215; Ginet 1990: 13–14). One possible reply is to say that the substance causes the effect to begin at a specific time *in virtue of having some relevant property*, although it is not the substance's having that property, but the substance alone that does the causing (considered by Clarke 1996: 201). The difficulty for those of us who do not reject non-object causes is that the event (or property instance or fact) of that substance's having that property will also be a cause of that effect, leaving no room for the efficacy of the substance per se. Those who advocate *only* object-causes to the exclusion of event/fact/property instance causation will not find this objection compelling, but that view is itself highly implausible.

1.5 The Value of a Variable View

One other view that should be mentioned briefly is that token causes and effects are actual values of variables and that token causal claims have the form, '*X* assuming some actual value on some particular occasion (for some particular individual) caused *Y* to assume some actual value on that occasion' (Woodward 2003: 74; see also Spirtes, Glymour, and Scheines 2000). For the hammering of a window to cause that window to shatter is for the actual value of the variable *H* {*hit, not hit*}, to cause the actual value of the variable *S* {*shatter, not shatter*}. The difficulty with trying to assess this view is that it is unclear what 'actual values of a variable' is supposed to pick out. One possibility is that they are determinates of the variables of which they are values (Woodward 2003: 39), but that would make them property types, say in the form of universals or classes of tropes, which are unsuitable as token causes/effects. On the other hand, the actual value of a variable might be an event or a fact or a property instance. Since I will assume that the ontological clarification of the 'value of a variable' view will put it in one of these camps I will not discuss it separately.

2. Unified or Pluralistic Theory of Causal Relata?

An account of causal relata is 'unified' if either it excludes irreducibly different kinds of causes/effects or posits causes/effects of irreducibly different types, but makes causal relations among all but one type derivative (Mellor 1995: 109). Unified theories of the first sort reject all but one kind of causal relata (for example, event-only accounts (Beebee 2004) or assert that various types of causal relata are, in fact, species of a common type (Menzies 1989). A unified account of the second sort does not require that there not be irreducibly different types of causal relata. There are two general arguments meant to establish that a

theory of causal relata *must* be unified, one based on the non-ambiguity of 'causes' and the other on the correlations of causal relations between facts and causal relations between particulars. Neither argument is compelling.

2.1 The Non-Ambiguity Argument

The non-ambiguity argument runs thus: if there were non-reducibly different kinds of causal relata and no one type of causal relata was such that the causal relations of all other types reduced to the causal relations of that type, then there would be different kinds of causal relations corresponding to these different relata. 'Causes', however, is not multiply ambiguous, picking out a wide array of different kinds of causal relations on different occasions. Hence, causal relata are unified in one of these senses (Menzies 1989: 62).

The truth of the claim that 'causes' is non-ambiguous is supposed to be obvious and evidenced by how we normally think and talk about causation (ibid. 65; see also Lewis 2004*b*: 286 for an argument against ambiguity based on ordinary thinking. But see Vendler 1967 who argues that 'cause', 'effect', and 'result' pick out different relations with varying relata).

My main worry about this argument is not with the non-ambiguity claim, but with the claim there could be no single causal relation that non-derivatively relates relata of irreducibly different types (Sanford 1985: 283). This claim needs some defence. One possible defence is to appeal to a more general principle that there are no relations with irreducibly different kinds of relata. But that general principle is not obviously true. Indeed, it would appear that there are relations that can hold among irreducibly different kinds of relata. For example, the relation of 'being distinct from' could hold between physical objects but also between irreducibly non-physical objects if there are any. Or, consider the relation 'being a part of'. That relation can hold between concrete physical objects, but arguably also between abstract objects: a sub-class is a part of a class (Lewis 1991). On the other hand, it might be argued that there is something special about the causal relation that rules out irreducibly different kinds of relata for this relation. However, determining whether or not there could be no single causal relation that relates irreducibly different kinds of relata must await an acceptable philosophical account of that relation. Furthermore, even if we grant that there is some evidence for the claim that 'causes' is not ambiguous, that is not the end of the argument. It may turn out that on deeper inspection of our causal practices, there are theoretical reasons for rejecting the assumption that there is only one kind of causation (see e.g. Hall (2004*b*: 253–4) who distinguishes two kinds of causation).

2.2 Mellor's 'Non-Independent' Argument

The second argument for the necessity of a unified theory runs thus: consider the causal pair, 'Don dies because he falls' and 'Don's fall causes his death' (Mellor 1995: 108). The causes/effects in the first case are facts, but in the second, particulars (events). These claims stand and fall together, but that would not be the case if there were independent causal links for particulars and facts (ibid. 130). Since the only causes/effects are facts and particulars, and facts and particulars are irreducibly different in kind, then one must derive from the other:

sentences reporting causal relations between one type reduce to sentences reporting causal relations between relata of the other type.

In the continuation of this argument, Mellor (ibid. 135) tries to show that fact-causation, not particular-causation, is non-derivative. He argues that in cases of causation by absence there are no suitable 'negative' particulars—such as negative events—to serve as causes/effects, but there are negative facts to play those roles. 'Since therefore causation cannot always link particulars, factual causes and effects cannot all be reduced to particular ones. The reduction, if any, will have to go the other way'.

This argument depends on the existence of absence causation. If absence causation is rejected, this argument does not go through. Absence causation will be covered in the context of our evaluation of the various candidates for causal relata which begins in sect. 3. As we shall see (sect. 3.3), there are good reasons for rejecting the existence of absence causation. The intrinsicness of causation is inconsistent with absence causation and, hence, Mellor's argument fails. (In sect. 3, I will also argue that there is no other convincing argument—independent of absence causation—*for* fact causation and, in sect. 4, that there is a good argument *against* the possibility of facts as causes based on certain spatial and temporal aspects of the relation between causes and effects.) In any case, the failure of these two arguments for a unified theory does not mean that the best theory is not unified.

3. THE CASE FOR FACTS AND AGAINST EVENTS AND PROPERTY INSTANCES

The case for fact-causation consists in citing causal phenomena that supposedly can be handled if causal relata are facts, but not otherwise. The three phenomena are overdetermination, iterated causation, and absence causation. As we shall see, none of these arguments is convincing.

3.1 Overdetermination

Some philosophers argue that overdetermination is incompatible with event views of causal relata, coarse-grained or otherwise, but not with the fact view (Bennett 1988: 140). This argument can also be directed against property instance theories of causal relata. The argument runs thus: suppose that Jones's death is overdeter-mined by two independently causally sufficient events, rifle firings f and f'. Neither of *these* events is individually a cause of the death, it is argued, either because neither one by itself makes a difference to the death (ibid.) or because neither raises the probability of the death (Mellor 1995: 102). And there are no other viable *event* candidates. In particular, there are difficulties with the most obvious candidate: the *disjunctive event* whose disjuncts are the overdeterminers. First, 'disjunctive' events with highly dissimilar disjuncts are objectionable, at least as causes, and there is no way to guarantee that the disjuncts of the disjunctive causes of overdetermination causation will not be highly dissimilar (Lewis 1986*a*: 267). Second, some philosophers go further and argue that sentences, propositions, and facts, but not events, are in the right categories for the operation of disjunction (Bennett 1988: 140; 1995: 41). On the other hand, there is a viable

fact candidate: it is the fact that at least one of the firings, f or f', occurred that caused the death. Worries that apply to disjunctive events do not apply to disjunctive facts: 'if we move from events to facts, all comes clear' (Bennett 1988: 140).

The event proponent can respond in a number of ways. He might claim that every aspect of the death event (and of any event) is essential to it—maximum fragility—and had one of f and f' failed to occur, that would have made *some* difference to when or how the death occurred, and so that very same death would not have occurred. There can be no such thing as overdetermination. This response, however, faces serious objections. First, maximum fragility gives counterintuitive results in some cases: slightly postponing the death of one's patient will count as a cause of that death, at least if counterfactual dependence between distinct events is sufficient for causation (Lewis 1986*a*: 250). Second, overdeterminers that individually make no difference to the time or manner of occurrence of the effect seem to be possible (Schaffer 2003*b*: 27).

A second response cites the event that is the spatially discontinuous mereological sum or fusion of f and f' as the cause of the death, not f and f' individually (Lewis 1986*b*: 212). This response will not be convincing to those who either reject or have unclear intuitions about the principle that for any set of events those events always have another event as their sum (Bennett 1988: 140). Furthermore, as Schaffer (2003*b*: 38) points out, if the individual overdeterminers are not efficacious, it is hard to see how the sum could be, and if they are, there is no need for the fusion to save the event view.

A third response, which I think is the most promising, involves arguing that the individual overdeterminers *are* causes of the death. Schaffer (2003*b*), for example, argues for that conclusion partly on the grounds that individual overdeterminers (a) play the typical roles of causes (for example, predicative and explanatory roles) and (b) are connected by complete processes to their effects. Any tendency to think otherwise is based on the mistaken assumptions that effects must counterfactually depend on their causes (or stand in chains of counterfacual dependence to their causes) or that causes must raise the probability of their effects. Pre-emption cases demonstrate the falsity of both of these assumptions.

3.2 Iterated Causation

In cases of 'iterated causation', an instance of the causal relation itself appears to be a cause or an effect; for example, suppose that the rock's impact causes the shattering of the window because the window is so fragile. The phrase 'the rock's impact causes the shattering of the window' seems to pick out the effect of the fragility; but an instance of the causal relation would seem not to be an event (or property instance). The effect is the *fact* that the rock's impact causes the shattering (Mellor 1995: 110; Needham 1988: 215–16; Kim [1976] 1993: 51).

Two alternative responses are available to the event proponent (analogous responses are open to the property instance proponent). First, deny that there is iterated causation. The fragility is a cause, along with the rock's impact, of the shattering. It does not cause the fact that the rock's impact causes the shattering (Kim [1976] 1993: 51). Alternatively, argue that

the 'fragilility' sentence can be rewritten as 'The fragility of the glass causes the rock's impact's causation of the shattering of the glass' so that both of the terms that flank 'causes' are referring terms picking out events (Noordhof 1998*a*: 857). One of these responses is likely to be right, and I suspect it is the first.

3.3 Absence Causation

The strongest argument for the fact view is based on causation by and of absences. The *absence of an event* (or property instance) rather than an *event* (or property instance), it seems, can be a cause/effect. If gardener Jones's failure to water the plant on Tuesday caused the death of the plant, it is the *non-existence* of any event (or property instance) of a certain type —a watering-of-the-plant-on-Tuesday-by-Jones type event—that is said to have a certain effect rather than an event (Bennett 1988: 140–1; Mellor 1995: 131–5; Lewis 2004*b*). On the other hand, absence causation can be accommodated if causes/effects are facts, at least if facts are true propositions: in this case, the cause is the fact that Jones did not water the plant (Bennett 1988: 140–1; 1995: 42; Mellor 1995: 134; Needham 1988: 215).

The events-only proponent may respond in one of four main ways (the property instance proponent has analogous options): (1) identify causally involved absences with purely negative events, (2) identify causally involved absences with positive events of some kind, (3) deny that there is causation by and of absences, or (4) accommodate absence causation in an event framework by asserting that causation is a non-binary relation (I will consider this response in sect. 7). A fifth option involves denying that in cases of absence causation, causation is a relation, but holding that when causation is a relation, it relates only events (Lewis [2000] 2004*a*: 100). The most promising of these strategies is the 'denial' approach.

The first strategy posits merely negative events 'which exist by definition just in case some corresponding positive events... do not exist' (Mellor 1995: 133). The failure of Jones to water the plant is a genuine event, but it is not identical to any positive event including whatever Jones did do or might have done in place of watering the plants. This strategy must contend with an argument put forward by Mellor:

Suppose instead that 'Don does not die' is made true by a single *negative event*, Don's survival, which exists just when Don is not dying. To make 'Don does not die' entail both 'Don does not die quickly' and 'Don does not die slowly', Don's survival will have to be both quick and slow; but it cannot be both, so it does not exist. (Mellor 1995: 133–4)

In short, positing negative events gives rise to absurdities and should be avoided. (This argument is disputed by Edgington (1997: 422) and defended by Menzies (1989: 66–7) and Persson (2002: 136).)

The second defence identifies a cause-absence with whatever non-disjunctive positive event (or property instance) happened in place of the absent event. The watchman's omitting precautions is identical with the nap he took at that time. But that cannot be right since (1) it may be true that had the watchman not taken the nap, he might still have omitted the

precautions, and this is prima facie evidence against identifying the napping event with the cause or the omission (Dowe 2000: 127), and (2), in Lewis's (2004*b*: 282) case of a void as cause, there is no positive event that happened in the place of the absent event.

Another version of the second defence identifies the causally involved absence with a disjunctive event, the disjuncts of which are actual and merely possible positive events. The watchman's omission is essentially the absence of precautions, an event that is essentially specifiable as a disjunction of all the actual and possible ways for the watchman to omit precautions—his napping-or-loafing-or-chatting-or-... (Lewis [1973] 1986*c*: 190). This defence may, however, require causes with 'highly varied disjuncts', whereas we generally reject claims such as 'John's talking-or-walking caused the spraining of his leg muscle' (ibid.). Furthermore, there is no suitable disjunctive event with which we can identify a void-cause (Lewis 2004*b*: 282).

A third line of defence is to deny that there is any absence causation. This is the defence I prefer, but there are different ways to motivate this defence. (1) Beebee (2004) argues that the denier of absence causation is no worse off with respect to common sense than is the typical supporter of absence causation. The denier rejects commonsensically acceptable claims of absence causation, but the typical defender accepts some absence causation claims rejected by common sense: 'the failure of Brown—who lives on the other side of the planet—to water the plant caused the plant to die'. Doing full justice to common sense—by distinguishing causally between the gardener's failure and Brown's failure to act, for example—will import into a metaphysical account of causation common sense's mistaken judgement that the moral status of an absence is relevant to its causal status (Beebee 2004).

Alternatively, (2) Dowe (2000: 125) argues that we have an 'intuition of difference' between absence 'causation' and non-absence causation such that if pressed, we do not think, for example, that the gardener's inaction is *literally* causally connected to the death of the plant. (Non-reflective common sense is taken in by the similarity between real causation and absence 'causation' (ibid. 135; for a reply to Dowe see Schaffer 2004).)

However, (3) the argument against absence causation that I find most convincing appeals to a feature of the causal relation, its intrinsicness relative to the laws of nature. (Lewis [2000] 2004*a*: 85) and Hall (2004*b*: 249–52) take this to *disfavour* the intrinsicness of causation (or for at least one kind of causation for Hall).) Duplicates of the same sequence of positive causes will exhibit the same causal relationships, including those of any absences, even if embedded in different extrinsic settings if the laws are kept constant (Hall 2004*b*: 250–1). But now consider a case that shows that if absences can be causes, causation is not intrinsic relative to the laws of nature: had Smith grabbed Jones's arm, Jones would not have succeeded in watering (*w*) the plant and it would have died. Suppose that Smith's failure to grab Jones's arm (*o*) is a cause of the survival of the plant (*s*). However, a duplicate of the *w*–*s* sequence embedded in a different extrinsic setting—say, in which Fred also watered the plant—would be such that the duplicate of *o*, *o*′, is not a cause of the survival, since *o*′ in that circumstance is irrelevant to the plant's survival. (This is based on a case in Hall 2004*b*: 249–2.)

If absences could be causes, then causation would not be an intrinsic relation relative to the laws of nature, but it is. Why is common sense misled in cases of absence causation? Common sense wrongly takes certain non-intrinsic relations such as counterfactual dependence and probability increase to be invariable signs of causation since they tend also to be present in

cases of genuine causation.

The rejection of absence causation has implications for our earlier discussion of Mellor's argument for fact-based, unified theory of causal relata (sect. 2.2). It means that Mellor's argument will not go through.

4. THE CASE AGAINST FACTS

But we can go further. Not only are the arguments *for* facts as causal relata ineffective, but there are strong reasons *against* facts as causal relata. These arguments appeal to certain features of causation that are incompatible with facts (true propositions) as causal relata, the most compelling of which relies on the spatial and temporal relations of causes and effects.

4.1 Interaction

One 'category mistake' objection to facts as causes runs thus: causes interact, exert force, and transmit energy, but facts/propositions do not. A true proposition, not being in the world but about the world, is 'categorically wrong for the role of a puller and shover and twister and bender' (Bennett 1988: 22). In response, Bennett asserts that this objection itself is guilty of a category mistake. Causes are *also* categorically wrong for these roles. Only *objects* interact, exert force, and transmit energy, and objects are not causes. Propositions and causes are *alike* in failing to interact. In response, I would suggest that this objection be restated as follows: even if causes per se do not transfer energy, (some) causes and effects stand in a relation to quantities of energy that propositions cannot. More specifically, at least some causes/effects can be redescribed as manifestations of transferred energy, even if the energy was transferred from the cause-object to the effect-object, but no proposition can be so redescribed. As Lewis ([2000] 2004*a*: 100) says, 'we distinguish between the cause itself and the true proposition that describes it'.

4.2 Spatial and Temporal Location

The strongest objection to facts as causal relata is that causes are dated and located particulars. Facts/propositions are not (Menzies 1989: 61, 74; Hausman 1998: 23). Propositions do not have, exist at or obtain at specific times or locations. The fact that an explosion occurred is not located at the time and place of the explosion, or anywhere else. Causes and effects, on the other hand, are dated/located, and they must be since causes and effects bear temporal and spatial relations to each other that require that they be dated particulars. For example, causes generally occur earlier than their effects and causes are generally spatially contiguous with their direct effects (even if one denies that temporal priority or contiguity are necessary features of causation). But no fact is earlier than or spatially contiguous with any other fact. The true proposition that the match was struck at 2 p.m. is not earlier than the true proposition that the building on Elm Street burned to the ground at 3 p.m. Furthermore, the efficacy of any token of a single type of event with respect

to any specific effect will depend on the time and location of that token. Not just any striking of a match at any time is a cause of this specific match lighting. Pairing up just *this* striking with *this* lighting will depend in part on when and where this striking and this lighting occurred or were realized or existed (Hausman 1998: 23; for a reply to this line of argument see Mellor 2004: 320).

4.3 The Slingshot Argument

There is also a logical argument against the claim that causes and effects are facts (Davidson [1967] 1980: 152–3). This argument purports to show that if some facts cause some facts, then all facts cause all facts, and since the latter is false, no facts cause any facts. The argument depends on two assumptions. (1) Substitution of co-

referring singular terms into causal contexts will not change their truth-value. If 'Elm Street' and 'Main Street' co-refer, then if c caused it to be the case that there was a fire on Elm Street, then c caused it to be the case that there was a fire on Main Street. (2) Substitution of logically equivalent sentences will not change the truth-values of causal sentences. For example, if the fact that (f_1 and f_2) caused it to be the case that f_3, then the fact that (f_2 and f_1) caused it to be the case that f_3.

One version of the argument runs thus: suppose that some facts are causally related and, in particular:

> (a) The fact that there was a short circuit caused it to be the case that there was a fire on Elm Street.

But now notice that 'there was a short circuit' is *logically equivalent* to '[the class of xs such that both (x is identical with x) and there was a short circuit] is identical to [the class of xs such that (x is identical with x)]' (Mackie 1974: 250). By (2) we can substitute the latter for 'there was a short circuit' into (a) to get:

> (b) The fact that [the class of xs such that both (x is identical with x) and there was a short circuit] is identical to [the class of xs such that x is identical with x] caused it to be the case that there was a fire on Elm Street.

In addition, 'the class of xs such that both (x is identical to x) and there was a short circuit' co-refers (to the set of everything) with 'the class of xs such that both (x is identical to x) and it rained in 2005' if 'there was a short circuit' and 'it rained in 2005' are both true. By (1), that means we can substitute the latter for the former in (b) to get:

> (c) The fact that [the class of xs such that both (x is identical with x) and it rained in 2005] is identical to [the class of xs such that x is identical with x] caused it to be the case that there was a fire on Elm Street.

Finally we can replace '[the class of *x*s such that both (*x* is identical with *x*) and it rained in 2005] is identical to [the class of *x*s such that *x* is identical with *x*]' with the logically equivalent 'it rained in 2005' to get:

(d) The fact that it rained in 2005 caused it to be the case that there was a fire on Elm Street.

A similar series of substitutions on the right side of (a) will give us:

(e) The fact that it rained in 2005 caused it to be the case that Bush was elected in 2004.

If any fact causes any fact, all facts cause all facts given these principles of substitution. Hence no fact causes any fact since not all facts cause all facts.

This and similar arguments, however, have been challenged. Here are some examples of those challenges.

(1) Cummins and Gottlieb (1972) reject the claimed logical equivalence of sentences such as 'there was a short circuit' and '[the class of *x*s such that both (*x* is identical with *x*) and there was a short circuit] is identical to [the class of *x*s such that (*x* is identical with *x*)]'. The latter, but not the former implies the existence of the universal set, at least on certain assumptions about the class abstraction operator. (If the existence of the universal set is guaranteed to exist by logic alone this criticism fails (Mackie 1974: 252).)

(2) Mackie (ibid. 253) rejects the *successive* application in the argument of the two principles of substitution. When 'the fact that there was a short circuit' is replaced with its logical equivalent in (b) a singular term is introduced—'[the class of *x*s such that both (*x* is identical with *x*) and there was a short circuit]'—in a position such that it is not replaceable *salva veritate* by a co-referring term (ibid. 254).

(3) Bennett (1988: 39–40) distinguishes different kinds of facts with varying patterns of substitutability. He argues that the kinds of facts that *causation* relates are not transparent with respect to any definite descriptions that are not being used 'merely as pointers to their referents', whereas the kinds of facts to which appeal is made in this argument are transparent in this way (see also Menzies 1989: 81–3). 'So we have the resources of expressing fact-causation statements without falling into the clutches of the Frege–Davidson argument' (Bennett 1988: 40).

(4) Sharvy and Anscombe raise scope-based criticisms that apply to this line of argument (Sharvy 1970; Anscombe 1969). Applying Sharvy's objections to causal contexts, we can say that co-referring definite descriptions with small scope are not intersubstitutable *salva veritate* in causal contexts, but those with large scope may be. (Large scope: 'Concerning the dog with blue eyes: there was a bite mark on Jones because it attacked him.' Small scope: 'There was a bite mark on Jones because the dog with blue eyes attacked him.') The mistake in the argument occurs when the co-referring description 'the class of *x*s such that both (*x* is identical to *x*) and it rained in 2005' is substituted for 'the class of *x*s such that both (*x* is identical to *x*) and there was a short circuit'. That substitution is illegitimate since these

descriptions appear with small scope.

(5) Neale denies that causal contexts obey the principle of the intersubstitutability of logical equivalents. For example, there is no good reason to think that 'Catiline fell because Tully denounced him' entails or is entailed by '(Catiline fell and (it is raining or it is not raining)) because Tully denounced him' (2001: 220). (For a related point see Menzies (1989: 83–4); Sanford (1985: 290) also rejects this principle of substitutability.)

(6) Another strategy is to admit that co-referring singular terms can be substituted *salva veritate* in causal contexts but argue that definite descriptions are not singular terms when properly analysed via Russell's Theory of Descriptions (Lowe 2002: 172; Neale 2001: 220).

(7) Mellor rejects the principle of the substitutability of co-referring terms into causal contexts on the grounds that we cannot go from the true causal statement, 'the fact that Don's rope is the weakest caused it to be the case that Don's fall is the first fall', to 'the fact that Don's rope is the weakest caused it to be the case that Don's fall is Don's fall' even though Don's fall is the first fall. The latter causal claim is certainly false since the necessary fact that Don's fall is Don's fall does not causally depend on anything. Hence, this principle of substitution is false (Mellor 1995: 117; for criticism of Mellor's argument see Rodriguez-Pereyra 1998).

Whatever one makes of the slingshot argument, we have sufficient grounds for dismissing facts as causal relata: some causes/effects are redescribable as manifestations of energy, but no facts are, and facts do not have specific spatio-temporal locations, but causes/effects do. We can move on to deciding between coarse-grained events and various fine-grained alternatives to them.

5. THE CASE FOR A FINE-GRAINED ALTERNATIVE TO COARSE-GRAINED EVENTS

In this section, I will consider two arguments for a fine-grained, non-fact alternative to coarse-grained events, one from the behaviour of causal sentences under emphasis and the other from the transitivity of the causal relation. The transitivity argument is the stronger of the two.

5.1 Emphasis

Dretske's argument from emphasis begins with the claim that the substitution of nominalizations of differently emphasized 'allomorphs' of the same proposition—say, 'Socrates *drank hemlock* at dusk' versus 'Socrates drank hemlock *at dusk*'—into the same causal context does not necessarily preserve the truth-value of the sentence. For example,

(1) Socrates' *drinking hemlock* at dusk causes his death

may be true but

(2) Socrates' drinking hemlock *at dusk* causes his death

may be false. Since causal contexts are referentially transparent, it follows that those nominalizations, 'Socrates' *drinking hemlock* at dusk' and 'Socrates' drinking hemlock *at dusk*' differently refer. They refer to different event aspects, not events, according to Dretske (1977). (On the other hand, Achinstein (1983) argues that since these expressions co-refer, causation is not a relation and there are no causal relata. Davidson ([1967] 1980: 161) considers a related argument based on sentences such as 'Socrates' death was caused, not by the fact that he drank hemlock at dusk, but by the fact that he drank hemlock', and he responds by classifying such sentences not as causal sentences, but as rudimentary causal explanations; but see Mellor's objection (1995: 130–1)).

The defender of coarse-grained events might respond to this argument by questioning the interpretation of the linguistic data on which the argument relies. Boer, in particular, argues that although speakers asserting (1) and (2) may differ in what they assert, they do so by asserting more than what (1) and (2) mean (1979: 294). Emphasis in causal contexts affects only what a speaker 'indirectly asserts', not what he directly asserts (ibid. 286). What is directly asserted does not differ in truth value and Dretske's argument cannot get started. Dretske must show why this alternative account of the linguistic data is wrong.

Furthermore, even if it were granted that this argument works against Davidsonian events, it does not succeed in showing that event aspects are the causal relata. It does not, for example, rule out Kimian events. The nominalizations cited in the argument can be interpreted to refer differently to different Kimian events. One can claim that the Kimian event with *drinking hemlock* as its constitutive property is picked out by 'Socrates' *drinking hemlock* at dusk' and it is the cause of his death, and that a non-identical Kimian event with a different constitutive property is picked out by 'Socrates' drinking hemlock *at dusk*' and it is not the cause of his death (Menzies (1989: 77); for a similar response see Kim (1977: 103). For a different line of criticism see (Ehring 1987*b*). The Lewisian might make a similar response but with the relevant nominalizations picking out different events with different essential properties.)

It is also worth noting, since I favour tropes as the causal relata, that the phenomenon of emphasis can be accounted for by tropes (which is not to say that Dretske intends to exclude tropes). One might take these nominalizations to pick out a trope-based version of Kimian events the constitutive properties of which are not universals but tropes, or, to pick out event aspects, but read as tropes.

5.2 The Transitivity Argument Against Coarse-Grained Events

What I take to be the best argument against coarse-grained causal relata appeals to the transitivity of the causal relation. Consider the following scenario on the assumption that causal relata are coarse-grained events: the event (call this *d*) of Davidson's putting potassium salts into the fireplace occurs just as Jenny puts a lighted match into the fireplace. Following *d* there is a purple fire in the fireplace (call the purple fire, *c*). The fire then causes Elvis's death (call the death, *e*).

Assume that putting potassium salts into the fireplace (*d*) is a cause of the purple fire (*c*). The fire is a cause of the death (*e*). Putting potassium salts in the fireplace, then, is a cause of

the death, given transitivity. But that is not correct. Intuitively, putting potassium salts into the fireplace has nothing to do with the death (Ehring 1987*a*; 1997; for similar examples see Paul [2000] 2004; Hausman 1992). On a coarse-grained view, there would appear to be a three-event chain running from *d* through *c* to *e* such that *d* causes *c* in virtue of *c*'s being purple and *c* causes *e* in virtue of *c*'s having a certain high temperature. Since the 'middle event' *c*, is not required under this view to be such that the same property of *c* that is relevant to *c*'s being caused by *d* is the same property of *c* that is relevant to *c*'s causing *e*, transitivity can fail.

A defender of the coarse-grained event view might (1) deny that putting potassium salts into the fireplace causes the fire, (2) argue that putting potassium salts into the fireplace *does* cause the death, contrary to intuition, (3) reject causal transitivity, or (4) argue that by treating causation as a non-binary relation, this case can be handled without abandoning the event view (I will discuss this view in sect. 7).

The first option is not plausible for all versions of this argument. Suppose that in the same case, the purple fire *also* causes Jones to notice a purple fire. In that event, adding the salts causes Jones to notice the purple fire by causing the purple fire, if causal relata are coarse-grained events. We cannot then go on to deny that adding the salts does not cause the purple fire when focusing on the death (Paul [2000] 2004: 210). The second option of affirming that adding salts causes the death requires that we explain away the intuition that it does not. One possible 'explaining away' explanation is to say that we are mistaking the failure of counterfactual dependence of the death on the salts for a failure of causation. The proponent of this option can cite the fact that counterfactual dependence is not in general necessary for causation, as is clear in cases of pre-emption (Lewis [2000] 2004*a*: 99). However, one might reasonably respond that in a short causal chain, without pre-emption, the failure of counterfactual dependence *is* a reliable indicator of a failure of causation.

The third option is to reject transitivity for causation. This response is defended by citing other cases in which causal transitivity seems to fail, and it requires a more extended discussion. There are various cases that are supposed to threaten transitivity even for short, strong-linked chains:

> (A) A man's finger is cut off in an accident, but it is surgically reattached in such a way that it functions as well as it would have without the accident. The accident causes the surgery and the surgery causes the state of the finger a year later, but the accident does not cause the state of the finger a year later (Kvart 1991).

> (B) Billy runs towards Suzy to stop her from throwing a water balloon at the neighbour's dog, but he trips and fails. Suzy notices none of this and throws the balloon. The dog yelps when he is hit with it. Billy's running causes him to trip and his tripping causes the dog to yelp because 'if he hadn't tripped, he would have stopped her from throwing and so the dog wouldn't have yelped' (Hall 2004*a*: 184). But his running does not cause the dog to yelp.

How might the transitivity argument against the coarse-grained event view be defended in light of these attacks on causal transitivity? (1) One can follow Lewis and affirm transitivity in these cases. One can argue that the first event in both cases causes the final event in the

sequence and that any tendency to think otherwise can be explained away. Focusing on the accident case, we can explain away this tendency by reference to the fact we wrongly think that delayers (including the accident) are not causes, the fact that accidents are not *generally* conducive to healthy fingers, and the fact that we wrongly think that counterfactual dependence is required for causation (Lewis [2000] 2004*a*: 98–9). (2) Alternatively, we can treat these cases differently. We can argue that the accident does cause the healthy finger (as in (1)), but that in the Suzy case, there is no test of transitivity since the tripping does *not* cause the yelping. The tripping prevents Billy from continuing to run which would have prevented the yelp ('double prevention'). The only reason for treating the tripping as a cause of the yelping is that the latter is counterfactually dependent on the former and that is not reason enough: there is no causal process connecting them (Hall 2004*a*: 184).

The approach I prefer is this: we can grant that transitivity fails in cases of these kinds (A and B) or remain neutral as to whether or not transitivity fails in these kinds of cases. However, we can argue that in short-chained cases like these the source of any possible failure of transitivity is the threat–saviour structure of those cases. In threat–saviour cases an event initiates a threat to the occurrence of an event *e* but also initiates a saviour event that cancels that threat (Hall 2004*a*: 184). But the purple fire case does not have that structure. So even if transitivity fails in all threat–saviour cases, it does not follow that we cannot appeal to a limited form of causal transitivity in presenting an argument against coarse-grained events based on the purple fire case.

If causal relata are *property instances* including aspects of events, then this case is not a counterexample to causal transitivity. The instantiation of the property *adding salts* causes the instantiation of the property *being purple* by the fire but not the instantiation of, say, the property *having a high temperature* by the fire. The latter instantiation causes the death, the former does not. (See Ehring 1997 and Paul [2000] 2004: 211. This will also work for the Kimian if Kimian events can be reasonably expanded to include the exemplification of a universal by an event (Maslen 2004: 354; see Paul [2000] 2004 for discussion).) This response works whether nor not property instances are exemplifications of universals or tropes (Ehring 1997). The fact proponent also has the option of 'dividing up' the middle relata and on that basis arguing that the question of transitivity does not arise in this case. Furthermore, the Kimian/Lewisian might distinguish two 'middle' events, one with the constitutive/essential property of *being on fire which is purple* and one with the constitutive/essential property of *being on fire*, and attempt to argue that the first is caused by adding salts, but that that same event does not cause the death, and that the second is not caused by adding salts, but that that same event does cause the death. (For the Lewisian version, see Maslen 2004: 354; Paul objects to the Kimian version of this approach in her [2000] 2004: 208.)

6. In Favour of Tropes over Exemplifications of Universals

We have now set aside facts as causal relata (since they are not spatio-temporally located) and Davidsonian coarse-grained events (in the light of the transitivity argument). If we also set aside Lewisian events as making events too fine-grained, we are left with the property instance camp, which I will restrict to exemplifications of universals or tropes. (If Kimian

events survive the transitivity argument, they may be assimilated to the exemplification-of-a-universal view for the purpose of this discussion.) Is there any reason to prefer either exemplifications of universals or tropes?

To answer this question, we must again bring into play the causal relation. An account of causal relata should be compatible with both the formal features of the causal relation and its substantial nature. I have already made reference to certain formal or semi-formal features of that relation, and I will now bring into play what I take to be the best theory of the substantial nature of the causal relation, specifically, a persistence theory. According to persistence theories, causes and at least their direct effects are connected by a process that involves something that persists, but they differ as to what it is that persists in causal processes. Transference theories are built around the notion of the transference of persisting quantities of energy or momentum from the cause-object to the effect-object (Aronson 1971). Salmon's account of causation (1984) brings into play persisting structures and Dowe's account (2000) invokes the world-line of a persisting object that carries a conserved quantity.

I have argued elsewhere (1997) that the best form of a persistence theory posits qualitative persistence as the singularist connection between causes and direct effects. This has implications for deciding between exemplifications of universals and tropes as causal relata. More specifically, the exemplification-of-a-universal view of causal relata will have trouble with the use of the notion of qualitative or property persistence as a basis for a theory of causation. If the proponent of exemplifications of universals understands property persistence as consisting in a series of causally connected exemplifications of the same universal, which seems likely, then such an account will be circular in the context of *using* the notion of property persistence as part of an account of causation. On the other hand, circularity is avoided if property persistence is analysed as trope persistence, where that is a matter of a trope existing wholly at each moment of its existence over a period of time and not a matter of causal relations between momentary tropes. The exemplification proponent does not have a similar option of analysing property persistence as the identity of an exemplification of a universal at one time with one at another time—existing wholly at both times—since the temporal components of such exemplifications will differ and, hence, the exemplifications themselves will be numerically distinct. In short, the best account of causation will make use of the notion of trope persistence and that gives us a reason for preferring tropes to exemplifications of universals as causal relata. (It is worth noting that if causally involved tropes are necessarily momentary, then on a trope view, causal relata will consist in such momentary tropes, but if trope-causes/effects involve enduring tropes, as I have postulated, a causal relatum will consist in a 'trope at a time', the existence or presence of an enduring trope at a time and location.)

7. THE NUMBER OF CAUSAL RELATA

I will now consider the question of how many relata the causal relation has. The standard view is that causation is a binary relation. This view is supported by the fact that many 'causal claims ... make no explicit reference to any contrasts' (Schaffer 2003*a*: 10). Some philosophers, however, argue that causation is a three- or four-place relation. Moderate

smoking did not cause cancer in Jones *simpliciter* but only relative to an alternative cause, no smoking, and not relative to heavy smoking (Maslen 2004: 341).

The first argument for a non-binary view appeals to the emphasis sensitivity of causal contexts. Suppose that citing Susan's *stealing* a bicycle as a cause brings into play a contrast with Susan's buying a bicycle, or the like, but that if the emphasis were on 'bicycle', a different contrast would be brought into play (Hitchcock 1996: 276). By accepting that supposition, we can explain why differently emphasized expressions in the cause position (or effect position) will not necessarily preserve truth values: different alternative causes (effects) are thereby invoked. Hence, we should accept that supposition.

This argument, however, will be convincing only if alternative accounts of causal emphasis can be shown to be wrong, including those based on various fine-grained accounts of causal relata and those that do not concede any shift in truth values across differently emphasized 'allomorphs' of the same causal sentence (sect. 5.1).

A second argument involves cases in which there is more than one alternative to the would-be cause. For example, in evaluating the claim 'moderate smoking caused Jones's cancer', there are at least two alternatives to moderate smoking: no smoking and heavy smoking. It is argued that the relevant causal claim is ambiguous—moderate smoking in contrast to heavy smoking did not cause Jones's cancer, but in contrast to no smoking, it did (Hitchcock 1996: 270–1; Maslen 2004: 343). This argument, however, can be resisted as follows. First, there is no sense in which this specific instance of (moderate) smoking did not cause the cancer: a causal process links the cancer to the moderate smoking. Second, *if* this statement is read as indeterminate or ambiguous that is because it is being read as a causal explanation, which, unlike causation, may be contrastive.

A third argument claims that a non-binary view makes it possible to reconcile certain pairs of apparently inconsistent causal claims that we accept. Suppose that Watson pushes a boulder off a cliff in a way meant to miss Holmes with the intention of preventing Moriarty from trying to kill Holmes with the boulder, but the boulder kills Holmes. We accept the apparently inconsistent claims that Watson's pushing the rock caused Holmes to die and that Holmes dies *despite* Watson's pushing the rock. However, if we interpret the claims as being made relative to different alternative causes there is no inconsistency: Watson's push is a positive cause of death relative to no Moriarty push, but it is not a positive cause relative to a Moriarty push (Hitchcock 1996: 274).

In response, one might reasonably claim that there is no sense in which Watson's push is not a cause of the death at the level of token-causation since these events are connected by a causal process. In so far as we are wrongly inclined to think that Watson's push is not a cause of the death, it is because we correctly accept the *general-level* claim that Watson-type events prevent Holmes-type deaths. (But see Hitchcock (ibid. 273) who argues that on a binary view if Watson's push is a cause at the token level of the death, but 'Watson-type pushes' prevent 'Holmes-type deaths', then there must be different species of causation operating at these different levels, which is objectionable. He argues that treating both token- and general-level causal claims as non-binary has the virtue of avoiding different species of causation.)

A fourth argument is predicated on the claim that the transitivity of causation fails in certain cases such as the purple fire case (sect. 5.2) and that those failures can be explained if

causation is 4-place relation on the assumption that transitivity fails if the contrast events relevant to the 'middle event' do not remain constant (Maslen 2004). Adding salts in contrast to adding some other substance causes the purple fire in contrast to a yellow fire, but the purple fire in contrast to cold logs is a cause of the death in contrast to survival. The contrast event for the middle event is not constant and, hence, transitivity fails but only because causation is a 4-place relation among coarse-grained events. In fact, this argument might be turned around to support the binary view. If causation is not binary, then causation is not transitive since transitive relations are defined only for binary relations, but causation is transitive. This argument would need to be supplemented with a treatment of the various cases that have been offered as violating transitivity (e.g. Lewis's defence of transitivity (sect. 5.2)). (There is a fifth argument for a non-binary view based on the claim that it enables a reconciliation of absence causation and a positive-event-only view of causal relata (Schaffer 2003*a*: 11).)

In short, none of these arguments for a non-binary account is sufficiently convincing to abandon the binary view, a view that fits much of how we think about token-causation. In any case, any move to a non-binary view should at least wait on a definitive verdict in current debates about whether or not the causal relation is transitive.

8. CAUSALLY RELEVANT PROPERTIES AND THE THEORY OF CAUSAL RELATA

A theory of the causally relevant properties of causes will specify what distinguishes the causally relevant and irrelevant properties exemplified by a cause with respect to a given effect, on the common assumption that causes are efficacious in virtue of the properties that they exemplify and that causes exemplify both causally relevant and irrelevant properties. (But note that this common assumption that causes are efficacious in virtue of the properties that they exemplify may be inconsistent with causal transitivity (Ehring 1997: 82–3).) The three main theories of causally relevant properties are: A property P of a cause c is causally relevant to an effect e of that cause just in case (1) (nomological accounts) the instantiation of P by c is nomologically sufficient for the occurrence of e (Fodor (1989) offers a nomological account as a sufficient condition for causal relevance), (2) (counterfactual accounts) e would not have occurred had c not had P (LePore and Loewer 1987; for a modified counterfactual account see Yablo 2003), (3) (essentialist accounts) P is an essential property of c (Brawn 1995).

Are these theories of causally relevant properties compatible with the major theories of causal relata? The *coarse-grained* event view is compatible with the nomological and counterfactual accounts, but not with the essentialist account unless it is assumed that the coarse-grained events have some essential properties. However, the latter assumption may not be consistent with the Quinean/later Davidsonian coarse-grained view of events. Spatio-temporally coinciding events are identical on the latter view, but it might be argued that in some cases they fail to be identical since they have different essential/accidental properties, a possibility opened up by assuming they have essential and accidental properties.

The *Lewisian* event account is compatible with the essentialist account and would *seem* to be compatible with the nomological and counterfactual accounts. However, it might turn out

on closer inspection, given the emphasis on the distinction between essential and accidental properties, that of these three accounts of causally relevant properties, the Lewisian account can only be coupled with the essentialist account.

The *Kimian* event account of causal relata is not compatible with any of these accounts *if* the latter apply to properties *exemplified by events*. None of a Kimian event's merely exemplified properties are eligible for being causally relevant to any of its effects. However, if these theories are interpreted as applying to the constitutive properties of Kimian events, the Kimian view is certainly compatible with the nomological account and may be compatible with counterfactual accounts, if a Kimian event could have occurred without its constitutive property or if the counterfactual account can be stated without this requirement. The essentialist account is compatible with Kimian events only if it is assumed that the constitutive properties of Kimian events are essential to those events.

On a *trope* model of causal relata, the most natural assumption is that tropes are not efficacious in virtue of higher-order tropes if they have any, but only, so to speak, in virtue of themselves. In that case, there will be no distinction between an account of what makes a trope a cause of *e* and what makes that trope causally relevant to *e*. There is no room for a theory of causally relevant properties as separate from a theory of causation. (If trope-causes are efficacious in virtue of property *types* under which they fall, things will be more complicated (Robb 1997; Noordhof 1998*b*; Ehring 2003).)

Finally, if causal relata are *facts*, then there is no room for causally relevant (or irrelevant) properties *of* causes and effects since true propositions, even if they have properties, would seem not to be efficacious of their higher-order properties. Neither do true propositions have constitutive properties on analogy with the constitutive properties of Kimian events. Hence it is unclear how to talk about causally relevant and irrelevant properties given this theory of causal relata.

FURTHER READING

A classic statement of a coarse-grained event view of causal relata can be found in Davidson (1967; 1969), along with a discussion of how events should be individuated. Kim (1973) provides perhaps the best development of a fine-grained event account according which causal relata are events, which, in turn, are exemplifications of universals. Lewis (1973; 1986*a*) defends an event account such that causal relata are events, but events have essential and accidental properties, and Lewis discusses this view in the context of his counterfactual theory of causation. For the event aspect view see Dretske (1977), which argues for this view based on the emphasis sensitivity of causal contexts. For an argument for the trope view of causal relata based both on transitivity and a persistence theory of causation, see Ehring (1997). Mackie (1974), Bennett (1988), and Mellor (1995) each provide important defences of facts as causal relata as well as detailed discussions of the event view. Menzies (1989) discusses a wide range of theories of causal relata along with an argument for the claim that a theory of causal relata should be unified.

REFERENCES

ACHINSTEIN, P. (1983). 'The Causal Relation', in *The Nature of Explanation*. Oxford: Oxford University Press, 193–217.

ANSCOMBE, G. E. M. (1969). 'Extensionality Reconsidered', *Journal of Philosophy* 66: 152–9.

ARMSTRONG, D. M. (1997). *A World of States of Affairs*. Cambridge: Cambridge University Press.

ARONSON, J. L. (1971). 'On the Grammar of "Cause"', *Synthese* 22: 414–30.

AUNE, B. (1977). *Reason and Action*. Dordrecht: D. Reidel.

BARWISE, J., and PERRY, J. (1983). *Situations and Attitudes*. Cambridge: Cambridge University Press.

BEEBEE, H. (2004). 'Causing and Nothingness', in Collins, Hall, and Paul (2004: 291–308).

BENNETT, J. (1988). *Events and Their Names*. Indianapolis: Hackett.

—— (1995). *The Act Itself*. Oxford: Oxford University Press.

BOER, S. (1979). 'Meaning and Contrastive Stress', *The Philosophical Review* 2: 263–98.

BRAWN, D. (1995). 'Causally Relevant Properties', *Philosophical Perspectives* 8: 447–75.

BROAD, C. D. (1952). 'Determinism, Indeterminism, and Libertarianism', *Ethics and the History of Philosophy: Selected Essays*. New York: Humanities Press, 195–217.

BYERLY, H. (1979). 'Substantial Causes and Nomic Determination', *Philosophy of Science* 46: 57–81.

CAMPBELL, K. (1990). *Abstract Particulars*. Oxford: Blackwell.

CHISHOLM, R. M. (1966). 'Freedom and Action', in K. Lehrer (ed.), *Freedom and Determinism*. New York: Random House, 11–44.

CLARKE, R. (1996). 'Agent Causation and Event Causation', *Philosophical Topics* 24: 19–48.

COLLINS, J., HALL, E. J., and PAUL, L. A. (eds.) (2004). *Causation and Counterfactuals*. Cambridge: MIT.

CUMMINS, R., and GOTTLIEB, D.(1972). 'On an Argument for Truth-Functionality', *American Philosophical Quarterly* 9: 265–9.

DAVIDSON, D. ([1967] 1980). 'Causal Relations', *Journal of Philosophy* 64: 691–703; repr. in his 1980, 149–62.

—— ([1969] 1980). 'The Individuation of Events', in N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*. Dordrecht: Reidel, 216–37; repr. in his 1980: 163–80.

—— (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

—— (1985). 'Reply to Quine on Events', in LePore and McLaughlin (1985: 172–6).

DOWE, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

DRETSKE, F. (1977). 'Referring to Events', *Midwest Studies in Philosophy* 2: 369–78.

EDGINGTON, D. (1997). 'Mellor on Chance and Causation', *British Journal for the Philosophy of Science* 48: 411–33.

EHRING, D. (1987a). 'Causal Relata'. *Synthese* 73: 319–28.

—— (1987b) 'Compound Emphasis and Causal Relata', *Analysis* 47: 209–13.

—— (1997). *Causation and Persistence: A Theory of Causation*. New York: Oxford University Press.

—— (2003). 'Part-Whole Physicalism and Mental Causation', *Synthese* 136: 359–88.

FODOR, J. (1989). 'Making Mind Matter More', *Philosophical Topics* 17: 59–79.

GINET, C. (1990). *On Action*. Cambridge: Cambridge University Press.

HALL, E. J. (2004*a*). 'Causation and the Price of Transitivity', in Collins, Hall, and Paul (2004: 181–204).

—— (2004*b*) 'Two Concepts of Causation', in Collins, Hall, and Paul (2004: 225–76).

HAUSMAN, D. (1992). 'Thresolds, Transitivity, Overdetermination and Events', *Analysis* 52: 159–63.

—— (1998). *The Asymmetry of Causation*. Cambridge: Cambridge University Press.

HITCHCOCK, C. (1996). 'Farewell to Binary Causation', *Canadian Journal of Philosophy* 26: 267–82.

HONDERICH, T. (1988). *Theory of Determinism: The Mind, Neuroscience, and Life-Hopes*. Oxford: Clarendon.

KIM, J. ([1973] 1993). 'Causation, Nomic Subsumption, and the Concept of an Event'. *Journal of Philosophy* 70: 217–36; reprinted in his 1993: 3–21.

—— ([1976] 1993). 'Events as Property Exemplifications', in Brand, M. and Walton (eds.), *Action Theory*. Dordrecht: Reidel, 159–77; repr. in his 1993: 33–52.

—— (1977). 'Causation, Emphasis, and Events', *Midwest Studies in Philosophy* 2: 100–3.

—— (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

KVART, I. (1991).'Transitivity and Preemption of Causal Impact', *Philosophical Studies* 64: 125–60.

LEPORE, E., and LOEWER, B. (1987). 'Mind Matters'. *Journal of Philosophy* 84: 630–42.

—— and MCLAUGHLIN, B. (eds.) (1985). *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell.

LEWIS, D. ([1973] 1986*c*). 'Causation'. *The Journal of Philosophy* 70: 556–67; repr. in his (1986*c*: 159–71).

—— (1986*a*). 'Events', in his 1986*c*: 241–70.

—— (1986*b*). 'Postscripts to "Causation"', in his (1986*c*: 172–213).

—— (1986*c*). *Philosophical Papers II*. New York: Oxford University Press.

—— (1991). *Parts of Classes*. Oxford: Blackwell.

—— ([2000] 2004*a*). 'Causation as Influence', *Journal of Philosophy* 97: 182–97; expanded version in Collins, Hall, and Paul (2004: 75–106).

—— (2004*b*). 'Void and Object', in Collins, Hall, and Paul (2004: 277–90).

LOWE, E. (2002). *A Survey of Metaphysics*. Oxford: Oxford University Press.

MACKIE, J. L. (1974). *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.

MASLEN, C. (2004). 'Causes, Contrasts, and the Nontransitivity of Causation', in Collins, Hall, and Paul (2004: 341–58).

MEINERTSEN, B. R.(2000). 'Events, Facts and Causation'. *Poznan Studies in the Philosophy of Sciences and the Humanities* 76: 145–81.

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

—— (2004). 'For Facts as Causes and Effects', in Collins, Hall, and Paul (2004: 309–57).

MENZIES, P. (1989). 'A Unified Account of Causal Relata', *Australasian Journal of Philosophy* 67: 59–83.

MOORE, M. (2005). 'Causal Relata', *Philosophia Pratica Universalis: Festschrift for Joachim Hruschka, Annual Review of Law and Ethics* 13: 589–641.

NEALE, S. (2001). *Facing Facts*. Oxford: Oxford University Press.

NEEDHAM, P. (1988). 'Causation: Relation or Connective?', *Dialectica* 42: 201–19.

NOORDHOF, P. (1998*a*). 'Critical Notice: Causation, Probability, and Chance: D. H. Mellor, The Facts of Causation', *Mind* 107: 855–77.

—— (1998*b*). 'Do Tropes Resolve the Problem of Mental Causation?', *Philosophical Quarterly* 48: 221–6.

PAUL, L. A. ([2000] 2004). 'Aspect Causation', *Journal of Philosophy* 97: 235–56; repr. in Collins, Hall, and Paul (2004: 205–24).

PERSSON, J. (2002). 'Cause, Effect and Fake Causation', *Synthese* 131: 129–43.

QUINE, W. V. (1985). 'Events and Reification', in LePore and McLaughlin (1985: 162–76).

REID, T. (1969). *Essays on the Active Powers of the Human Mind*. Cambridge: MIT.

ROBB, D. (1997). 'The Properties of Mental Causation', *Philosophical Quarterly* 47: 178–94.

RODRIGUEZ-PEREYRA, G. (1998). 'Mellor's Facts and Chances of Causation', *Analysis* 58: 175–81.

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

SANFORD, D. (1985). 'Causal Relata', in LePore and McLaughlin (1985: 282–93).

SCHAFFER, J. (2003*a*). 'The Metaphysics of Causation', *Stanford Encyclopedia of Philosophy* (Spring 2003), ed. Edward N. Zalta, http://plato.stanford.edu/archives/spr2003/entries/causation-metaphysics/>, accessed 10 March 2009.

—— (2003*b*). 'Overdetermining Causes', *Philosophical Studies* 114: 2–45.

—— (2004). 'Causes Need Not Be Physically Connected to Their Effects: The Case for Negative Causation', in C. Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell, 197–216.

SHARVY, R. (1970). 'Truth-Functionality and Referential Opacity', *Philosophical Studies* 21: 5–9.

SPIRTES, P., GLYMOUR, C., and SCHEINES, R. (2000). *Causation, Prediction, and Search*. Cambridge: MIT.

SWINBURNE, R. (2000). 'The Irreducibility of Causation', *Dialectica* 51: 79–92.

TAYLOR, R. (1966). *Action and Purpose*. Englewood Cliffs: Prentice Hall.

VENDLER, Z. (1967). 'Facts and Events', *Linguistics in Philosophy*. Ithaca: Cornell University Press, 122–46.

WILLIAMS, D. C. (1953). 'On the Elements of Being', *Review of Metaphysics* 7: 3–18, 171–92; repr. in his *The Principles of Empirical Realism.* Springfield, Ill.: Charles Thomas, 1966.

WOODWARD, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

YABLO, S. (2003). 'Causal Relevance', *Philosophical Issues* 13: 316–28.

CHAPTER 20
THE TIME-ASYMMETRY OF CAUSATION

HUW PRICE BRAD WESLAKE

1. INTRODUCTION

One of the most striking features of causation is that causes typically *precede* their effects —the causal arrow seems strongly aligned with the temporal arrow, as it were. *Why* should this be so? This is the puzzle of the time-asymmetry of causation. In this chapter we offer an opinionated guide to this problem, and to the solutions currently on offer.

1.1 Hume's Semantic Conventionalism

A good place to start is with the parsimonious patriarch of philosophy of causation in the modern era, David Hume. Early in the *Treatise,* Hume offers this 'definition' of 'the relation of cause and effect': 'We may define a CAUSE to be "An object precedent and contiguous to another, and where all the objects resembling the former are plac'd in like relations of precedency and contiguity to those objects that resemble the latter" ' (*Treatise* 1. 3. 14). This proposal makes it a matter (literally) of definition that causes precede their effects. Hume takes the core of the causal relation to be the symmetric notions of contiguity and regularity, and proposes that we impose an asymmetry upon these symmetric relations, by labelling as 'cause' and 'effect' the earlier and later of a pair of appropriately related events. If Hume is right, then the relation between the causal arrow and the temporal arrow is merely a matter of *semantic convention*.

Hume's proposal has some evident attractions. It implies that there is no separate problem about the causal asymmetry, which is just an oblique way of referring to the temporal asymmetry. But despite its economical advantages, Hume's view has not been popular. There are two main objections. The first is that Hume's view makes the connection between causal asymmetry and temporal asymmetry *too tight*. Many philosophers have felt that there is an interesting issue as to whether there are, or could be, instances of *simultaneous causation,* in which the cause happens at the same moment as the effect; or even *backward* (or *retro-*) *causation,* in which the cause happens later than the effect. Hume's view turns these issues

into conceptual confusions.[1] If we share the intuition that backward and simultaneous causation are not obviously absurd, we must reject Hume's view, at least in its simple form.

The second difficulty with Hume's view is that it is *too weak,* in the following sense. Causation seems connected to *deliberation.* In particular, the temporal asymmetry of causation seems to have something to do with the fact that it doesn't make sense to deliberate with *past* ends in view. Hume's proposal does not begin to explain this fact. To see this, imagine that we have a ticket in a lottery drawn yesterday. The results have not yet been announced, and we are hoping that we have won. Why does it seem so absurd to try to do something *now* to ensure, or make it more likely, that our ticket was drawn from the barrel some hours ago? If Hume is right, it is no answer to be told that because the draw took place in the past, its outcome cannot be an *effect* of a present action. For on Hume's view, this just amounts to repeating the claim we were trying to *explain,* namely, that we act for later ends (i.e. for Hume, 'effects'), but not earlier ends ('causes'). If there were a present action that would *guarantee* our success in yesterday's draw, why should we care whether it could properly be said to cause it?[2]

The limitations of Hume's view thus bring into focus two general *desiderata* for an adequate account of the time-asymmetry of causation. It should explain the fact that the causal arrow is typically—though perhaps not *necessarily*—aligned with the temporal arrow. And it should help us to make sense of a matter of great practical importance in our lives, the fact that we can act for future ends but not past ends (at least in normal circumstances).

We will be stressing the latter point, in particular, at various stages in this chapter—we will call it the Practical Relevance Constraint (PRC). It turns on the intuition that an account of the time-asymmetry of causation should be able to explain the time-asymmetry of deliberation, or at least emerge as part of the same package. We shall argue that its ramifications are wider than usually appreciated; it creates difficulties for some popular attempts to explain the asymmetry of causation.

1.2 The Physicalist Constraint

Another constraint stems from physicalism—for example, from the intuition that the abilities the world grants us, and the restrictions it imposes on us, are determined ultimately by physics. Hence, apparently, we should look to physics for the origins and nature of the causal asymmetry. Yet this raises a new puzzle. Fundamental physics seems to be time-symmetric, in the sense that if it permits a process to occur in one temporal direction, it also allows it to occur in the opposite temporal direction. How could time-symmetric physics yield something as time-asymmetric as the cause–effect distinction?

One tempting response is to appeal to those parts of physics that are *not* time-symmetric, such as thermodynamics. We shall return to this approach below. First, it should be noted that some writers conclude at this point that there is no time-asymmetric causal arrow. A common view among physicists is that the only physically respectable notion of causation is time-symmetric: namely, the notion of what may be deduced from what in accordance with deterministic laws. For example, Stephen Hawking (1994: 346) describes his encounter with Reichenbach's (1956) work on the direction of time:

It laid great stress on causation, in distinguishing the forward direction of time from the backward direction. But in physics we believe that there are laws that determine the evolution of the universe uniquely. So if state A evolved into state B, one could say that A caused B. But one could equally well look at it in the other direction of time, and say that B caused A. So causality does not define a direction of time.

Clearly, this symmetric attitude does not explain the asymmetry of practical reasoning. Nor, apparently, is it consistently applied in science. Physicists use ordinary asymmetric causal reasoning as much as anyone else does, for example, in thinking about the consequences of possible experimental interventions.

A simple example: imagine a photon passing through two polarizers, on its journey from a distant light source. Consider the photon in the region between the two polarizers. Physicists, as much as anyone else, find it natural that the state of the photon at that point depends on the orientation of the first polarizer—the one through which it passed *in the past*. They find it highly counterintuitive that it might similarly depend on the orientation of the second polarizer—the one through which it passes *in the future*. This asymmetry is reflected in the description of such a case in textbook quantum mechanics, according to which the state of the photon reflects the fact that it *has passed* the earlier polarizer, but not the fact that it *will pass* the local future polarizer.

It is not only physicists who have taken the time-symmetry of fundamental physics to provide a reason for denying that there is any such thing as time-directed causality. This was also a motivation for the twentieth century's most famous philosophical critic of causality, Bertrand Russell (1912–13).[3] Again, however, Russell's view leaves us with a puzzle. What are we to make of the fact that we seem unable to influence the past? If Russell were right that 'physics has ceased to look for causes', would we be free to make money on yesterday's horse race? On the contrary, obviously, our puzzle would be intact and unsolved, as the issue as to why our practical abilites are so strongly aligned with the temporal arrow.

1.3 Hyperrealism

We might be tempted to respond to the tension between the time-symmetry of physics and asymmetry of causal dependence by denying physicalism—by regarding causation as something 'over and above' physics. Physics itself may be time-symmetric, but perhaps there is a further, causal, aspect of reality that is asymmetric. Call this the *hyperrealist* view of causation. It takes causation to be as real as the aspects of the world with which physics is immediately concerned, but not reducible to or supervenient on those aspects.[4]

The main difficulty with hyperrealism is that in putting causation beyond physics, it threatens to make it both *epistemologically inaccessible* and *practically irrelevant*. After all, if the causal direction is detached from physics, then presumably the world could have had the same physics, with an oppositely directed causal arrow—in which case, apparently, we have no way of knowing whether our ordinary ascriptions of the terms 'cause' and 'effect' are correct or back to front. Perhaps the past actually depends on the future. How could we tell?

And either way, what practical difference does it make to the choices we face as agents?[5] Hyperrealism thus seems an unpromising solution to the puzzle of the time-asymmetry of causation.

1.4 Grounding the Causal Arrow

Let's review the problem. It may seem that any explanation of the time-asymmetry of causation will need to rest on some account of the nature of causal asymmetry itself—that is, of the intrinsic *difference* between cause and effect (with the issue of time orientation set aside). But Hume shows us another possibility. Perhaps there is no causal asymmetry, as such —no asymmetric causal relation in the world—but only a semantic convention to label symmetric relations with an image of the past–future asymmetry.

By way of analogy, imagine someone puzzled by the difference between royalty and the rest of us. What (he wonders) are the distinctive qualities of *royal* individuals, and why are those qualities correlated with constitutional role—why are they found in particular among the families of hereditary rulers? The analogue of Hume's view—uncontroversial, presumably, in this case!—is that there are no such distinctive qualities. 'Royal' is simply a label applied by convention to the families of rulers of this sort, and the only asymmetry is the constitutional one.

At the opposite extreme from Hume lies hyperrealism. This view not only postulates a real causal asymmetry in the world, but takes it to be a primitive feature, not reducible to physics. We have seen that both extremes seem unsatisfactory. Among other failings, neither meets PRC—on both views, the practical asymmetry of deliberation remains mysterious.

At this point, there are two main options. The first agrees with the hyperrealist that there is a real causal asymmetry, but seeks to make it physical rather than 'extra-physical'. In other words, it seeks a physical asymmetry with the right relation to the temporal arrow—usually but perhaps not necessarily aligned past-to-future—and the right kind of relevance to our deliberative lives. Following Price (1996: ch. 6), let us call such an asymmetry a *third arrow*. It would provide a link between the causal arrow, on one side, and the temporal arrow, on the other.

If we could find a suitable third arrow, the following kind of account would be on offer:

1. The cause–effect distinction turns on the fact that causes are 'upstream' and effects 'downstream', with respect to the third arrow.
2. The link between the causal asymmetry and the temporal arrow turns on the fact that the third arrow has a prevailing temporal orientation; usually, though perhaps contingently, it points 'past-to-future'.
3. The relevance of the third arrow to deliberation ensures that this, too, picks up the usual temporal orientation of the third arrow itself.

Where might we find such a useful piece of philosophical weaponry? Not in (time-symmetric) fundamental physics, presumably, but this leaves the possibility mentioned earlier. The third arrow might be linked to some striking respects in which physics is *not* time-symmetric, such as the time-asymmetry of thermodynamic phenomena. We turn to this

proposal in a moment.

The second option is to side with Hume rather than the hyperrealist on the issue as to whether there is an objective causal asymmetry in nature. Perhaps Hume was right to deny this, though wrong in his alternative suggestion concerning the meaning of 'cause' and 'effect'. Hume proposed that these terms indicate the time-ordering of pairs of events in the appropriate (symmetric) relationship, but perhaps this misses the crucial point. In some cases, an asymmetry is a product of an asymmetric viewpoint on a symmetric state of affairs. Think of the distinction between the *left* side of the street and the *right* side; between *nearby* places and *remote* places; or between *locals* and *foreigners*. All these distinctions are drawn 'from a perspective' (and reverse their directions, in the obvious ways, if the perspective changes). As we shall explain, the main alternative to the third arrow strategy proposes that the direction of causation is a case of this kind; and that it is our perspective as *deliberators* that underpins the distinction between cause and effect.[6] We shall return to this proposal in due course.

2. THE SEARCH FOR THE THIRD ARROW

The most prominent example of the third arrow strategy is that of David Lewis. Though not originally proposed as an account of the causal asymmetry in terms of the thermodynamic asymmetry, Lewis's view turns out to be best defended along these lines. We shall explain why this is so, and then turn to a recent proposal in which the link is explicit.

2.1 The Asymmetry of Counterfactual Dependence

Famously, Lewis (1973) defends a counterfactual analysis of causation.[7] The central idea behind such an analysis is that it is typically the case for causally related events that had the cause not occurred, the effect would not have occurred. Of course it is also typically the case that had the effect not occurred, it would have been because the cause did not occur. So what the analysis requires in order to distinguish causes from effects is an analysis of a variety of counterfactual dependence according to which effects counterfactually depend on their causes but not vice versa. For our purposes such an analysis is also required in order to address the puzzle of the connection between the asymmetry of causation and the time-symmetry of physics. That is, the analysis should make it clear not only how it is that effects depend on their causes but not vice versa, but how this asymmetry is grounded in some asymmetric fact about our world consistent with the time symmetry of fundamental physics.

Lewis (1979) provided just such an attempt within the framework of his possible worlds analysis of counterfactuals.[8] According to this analysis, a counterfactual is true just in case, among worlds in which the antecedent is true, the consequent is true in at least one world closer to the actual world than any in which it is false. The analysis therefore requires an account of *closeness,* or *similarity,* between possible worlds. Lewis rejected the option of making the similarity relation one according to which by definition, for any possible world, worlds preserving the past are always more similar overall than worlds not preserving the past.

This would have turned the counterfactual account of causal asymmetry into a variant of Hume's conventionalism, which Lewis rejected for the first reason we discussed in sect. 1.[9] Instead, the similarity relation Lewis opted for was designed to make it a contingent matter that at least generally, with respect to the *actual* world, worlds preserving the past are more similar overall than worlds not preserving the past.

The contingent feature of the world supposed to secure this outcome can be understood by considering how the nearest possible world where the antecedent is true is to be identified, according to Lewis. Call the actual world w_0, the nearest world w_1, and the time of the antecedent t. Under the assumption of determinism, according to which two possible worlds are qualitatively identical either always or never, we know that if the past of w_1 is identical to w_0 and yet w_0 is different from w_1 at t, some violation of the laws of nature of w_0 must occur in w_1. This difference between the laws of nature in the two worlds Lewis refers to as a miracle. Intuitively, in w_1 the past is identical to w_0 up until just before t, at which point things go just differently enough to have the antecedent occur at t. What happens later is left to the laws of w_1 to settle.

Consider now w_2, a competitor to w_1 for similarity. We attempt to construct w_2 by following the temporally reversed strategy—in w_2 the future is identical to w_0 except for just after t, at which point things go just differently enough to have the antecedent occur at t. What happens earlier is left to the laws of w_2 to settle. To put it figuratively, in w_1 we run the tape forwards and diverge just in time to secure an alternative future in which the antecedent occurs, while in w_2 we run the tape backwards and diverge just in time to secure an alternative past in which the antecedent occurs.

What Lewis required here was a reason for thinking there was an asymmetry between w_1 and w_2 with respect to the actual world. His strategy was essentially to *deny* that there are worlds such as w_2, in which the antecedent world differs from the actual future only by a *small* miracle.[10] Lewis did not take this alleged *asymmetry of miracles* to be primitive; rather, he took it to reflect a contingent empirical asymmetry that he called the *asymmetry of overdetermination*. A *determinant* is defined by Lewis (1979: 474) as 'a minimal set of conditions jointly sufficient, given the laws of nature, for the fact in question', and what Lewis claims is that there are in our world many more future determinants than past determinants for events. Since we are assuming determinism, this is in addition to whole states of the world determining earlier times—as Lewis (1986*b*: 57–8) puts it, there are 'plenty of very incomplete cross sections that postdetermine incomplete cross sections at earlier times'. And so, if we believe that had some cause not occurred, the effect would not have, Lewis claims that even under the assumption of determinism we cannot conclude that if the effect had not occurred the cause would not have—since there generally exists some other (future) effect (or set of effects) sufficient, given the laws, to determine the cause. Figuratively, when we run the tape backwards and try to diverge just in time to secure the antecedent, we find that we cannot, since the antecedent is determined by many widespread facts about the future.

2.2 Overdetermination and Thermodynamics

Lewis himself professed uncertainty about the relationship between the asymmetry of overdetermination and that of thermodynamics. His paper ends with the remark: 'I regret that I do not know how to connect the several asymmetries I have discussed and the famous asymmetry of entropy' (Lewis 1986*a*: 51). However, he believed originally that the asymmetry of overdetermination is not a statistical asymmetry; and therefore, by implication, that it is distinct from the thermodynamic asymmetry, to the extent that the latter does rest on a statistical asymmetry. Field (2003: 458) reports that Lewis changed his mind about this, and came to regard the asymmetry of overdetermination as a statistical asymmetry. And an argument due to Elga (2000) makes it very clear that the asymmetry of over-determination is defensible, if at all, only in this form. Unless we restrict the options in the way that the second law of thermodynamics does, miraculous convergence is ridiculously easy.

Elga's argument exploits a very fundamental feature of a widely accepted statistical explanation of the second law of thermodynamics, the essential elements of which are due to Ludwig Boltzmann (1844–1906). Boltzmann's explanation combines two main ingredients. The first is a statistical consideration. For any macro-state of a physical system which is not already in thermodynamic equilibrium, there are many more microstates compatible with that macrostate whose evolution would be *towards* equilibrium, than microstates which would evolve away from equilibrium. This might seem sufficient to explain the fact that, in our experience, isolated systems do evolve towards equilibrium.

The flaw in this reasoning was first pointed out by Boltzmann's teacher and colleague, Josef Loschmidt (1821–95). The statistical considerations are time-symmetric. If they alone imply that entropy increases towards the future, then they alone would also imply that entropy increases towards the past: time-symmetric statistics cannot break the symmetry, to explain the monotonic increase of entropy we actually observe. To explain what we observe, we need to supplement Boltzmann's statistics with a second assumption, a time-asymmetric 'boundary condition'. We need to assume that the observed universe begins in an extremely low entropy condition, at some point in the distant past. Borrowing a term from Feynman (1965: 110), Albert (2001) calls this assumption the Past Hypothesis (PH).[11]

Loschmidt's point implies that the *actual* microstate of our familiar universe is always remarkably 'special', in the following sense. The vast majority of microstates compatible with the actual macrostate are associated with histories very *unlike* that of the actual world (as we believe it to be)—histories in which entropy *increases* towards the past, rather than *decreasing* towards the past. As Elga points out, this means that there is actually a huge *superabundance* of microscopic miracles, providing exactly the cases Lewis's asymmetry of overdetermination is meant to exclude: worlds that converge from very different histories, to differ from the actual world by a tiny local miracle. Without the restriction imposed by PH, in other words, the asymmetry of overdetermination would fail on an absolutely massive scale.

Elga's argument suggests that to the extent that there is an objective physical asymmetry of the kind that Lewis took to ground the asymmetry of counterfactual dependence, it involves macroscopic, statistical phenomena, of the same kind as ordinary manifestations of the thermodynamic asymmetry; dependent, in particular, on the same initial conditions.[12] Indeed, it is tempting to characterize these phenomena, generically, as examples of the dispersal of precisely the kind of macroscopic concentrations of energy that are produced by PH. To the

extent that Lewis's intuitions lead us in the direction of a genuine physical asymmetry—a possible candidate for a third arrow—it seems to be this one.

In a moment we turn to an explicit proposal for linking the asymmetry of causal and counterfactual reasoning to PH, from recent work by Albert, Kutach, and Loewer. Before that, let's distinguish two questions that need to be raised about Lewis's proposal. First, has Lewis successfully identified an objective temporal asymmetry with the right distribution to provide a third arrow—has he found a physical asymmetry in more or less the right place? Second, can the resulting account meet PRC—can it account for the asymmetry of deliberation?

We shall return to the latter question in sects. 3.3 and 3.4 below. Concerning the former, there are some evident difficulties. As Price (1996: ch. 6) notes, grounding causal asymmetry on a macroscopic statistical asymmetry seems likely to imply that there is no causal asymmetry at a microscopic or substatistical level. True, it is easy to impose an asymmetry at that level by fiat, by using the macroscopic asymmetry as a kind of 'signpost'. But this is much the same as Hume's view, with the reference to earlier and later replaced by reference to the direction in which entropy increases, or something similar. As a result, the same objections apply. Don't we exclude microscopic retrocausality by fiat, for example?

3. APPEALING TO THE PAST HYPOTHESIS?

The most explicit attempt to link the asymmetry of causation and counterfactual dependence to that of thermodynamics lies in recent work by Albert (2001), Kutach (2001; 2002; 2007), and Loewer (2007). For present purposes we ignore various differences between these authors, referring to the proposal collectively as the AKL view. (Note that Kutach himself no longer subscribes to the AKL view. See Kutach (forthcoming) for his current proposal, which is closer in spirit to the viewpoint that we endorse below.)

The AKL proposal tries to use PH to explain the asymmetry of counterfactual dependence. The basic idea is to argue that in virtue of PH, small, local changes—the kind of things we could use as 'causal handles', as Albert (2001: 128) puts it—produce much bigger and more diverse changes in the future than they do in the past. Intuitively, PH is supposed to do the job of ensuring that if we wiggle a causal handle in the present, we produce corresponding wiggles in the future but not in the past—or at least, not in the macroscopic, noticeable past. Loewer explains this idea using the figure of a tree, branching to the future but confined to one trunk in the past. PH is supposed to do the job of excluding (macroscopic) branching to the past. The initial plausibility of this idea is easily seen by recalling Elga's objection to Lewis's asymmetry of overdetermination. Elga's demonstration that convergence to the actual world is, *pace* Lewis, actually very easy, relies precisely on Loschmidt's anti-thermodynamic worlds —worlds *without* PH, in other words.

3.1 A Web Not a Tree

The AKL proposal has been sharply criticised in a series of papers by Mathias Frisch (2005*a*; 2007; forthcoming). In particular, Frisch challenges the claim that PH supplies the required tree structure. In many cases, he argues, the actual structure seems more like a web

than a tree. In other words, it contains divergence to the past, as well as the future—which would imply, by AKL's lights, that small, local changes could produce macroscopic changes in the past, as well as the future. For example, Frisch considers a gas in a two-chamber container, which was initially in one of two low-entropy conditions: all the gas was in the left chamber, or all the gas was in the right chamber. After the gas has dispersed between the two chambers, it may well be the case that only tiny local changes separate microstates evolved from the two distinct initial conditions. In this case, the AKL approach seems to imply that a tiny present change could cause the gas to *have been* in one chamber rather than the other. (As Frisch points out, thermodynamics itself implies that this kind of case is likely to be very common, for it is simply a consequence of equilibration.) Frisch also notes that even setting aside this kind of gross counterexample, the AKL approach seems unsatisfactory. The intuitive asymmetries of causation and counterfactual reasoning seem sharper, more general, and far less sensitive to the micro-macro distinction, than the AKL proposal can possibly account for.

3.2 Would a Future Hypothesis Prevent us Affecting the Future?

Another class of objections to the AKL approach rests on the observation that if it were true that PH (in conjuction with the time-symmetric resources noted above) were sufficient to explain our inability to affect the past, then—by symmetry of reasoning—an analogous low-entropy boundary condition in the future would prevent us from affecting the future. But would a 'Future Hypothesis' (FH) have this consequence? We think not.

The first question is whether such a future constraint would imply that our deliberative phenomenology would be a future-directed analogue of what we are trying to explain with respect to the past: the sheer apparent absurdity, at least in ordinary cases, of acting so as to influence the past. It is hard to see why this should be so. Restrictions in the distant future—even extreme restrictions, much tighter than PH itself—seem to have virtually no bearing on our present sense that we can affect the future. Suppose God tells us that as a matter of law, the final state, some fifteen billion years from now, will be constrained within some tiny region of phase space (comparable in size to that required by PH). Better still, suppose he offers to tell us the *actual* final microstate, to as many decimal places as we wish. Either way, the AKL tree of possible trajectories suffers the kind of pruning towards the future that PH requires towards the past. Do we lapse into fatalism, coming to think it absurd that we might seek to influence our immediate future? It is hard to see why we would, or should.[13] Hence, by symmetry, it is hard to see why a remote past hypothesis should be incompatible with taking ourselves to be able to affect the near past.

It might be objected that this argument trades too much on the fact that it considers only a *distant* future constraint. Setting aside the obvious reply that PH is rather distant too, let us turn to consequences of much closer future constraints. Would these necessarily be perceived as making deliberation absurd? On the contrary, we think, they might provide a new degree of control, an influence over matters previously thought to be independent of our actions.

To adapt an old example from the decision theory literature (Gibbard and Harper 1978: 136), suppose we believe that we are destined to meet Death at noon on a certain day. We

regard this as a lawlike future boundary condition.[14] It is now 09:05 on the fateful morning, and we sit in Aleppo airport, with a boarding pass for the flight to Damascus. We know that Death will meet us in one place or other; and moreover (since he refuses to fly) that he is already on the road to whichever place it is to be. Is it *absurd* to think that we are still free to choose whether to board the plane? On the contrary, apparently. While the boundary condition certainly deprives us of many options—the option to be anywhere other than Damascus or Aleppo at noon, for example, or to be anywhere at all, later in the day—it also yields some new abilities: in particular, the ability to influence Death's movements, even somewhat *earlier* on the day in question.[15]

The example suggests that while a lawlike future constraint can limit the options, it does not make it absurd to think that we exercise control within those limits. Within those limits, its effect seems to be not to prevent us from achieving ends, but to ensure that the world conspires to bring about those ends. Far from preventing us from achieving the ends, in other words, it gives us a new kind of control over *other* events—the ones that need to be appropriately arranged, in the light of the new constraint, for our ends to come to pass. This means, in particular, that we may be able to affect the remote present, and the past, via a kind of zigzag. We choose the future in some respects, and the future constraint ensures that the remote present and past keep in sync, in order to achieve the required final state. If this is how things would go in a world with lawlike future boundary conditions, shouldn't PH have the same kind of effect? Shouldn't it merely *limit* our capacities to influence the past, and compensate by giving us new powers—powers, say, to affect the remote present, by affecting bits of the past with which the antecedents of the remote present are necessarily correlated?

This possibility has been missed, apparently, because AKL have failed to notice an ambiguity in the requirement that we consider the consequences of small, *local* changes—causal handles, to use Albert's term. The requirement that the handles be local is needed to avoid a trivial falsification of the theory, because in the assumed context of a deterministic theory, it is immediate that large-scale differences will make a difference at earlier times, as well at later times. But this restriction to small, local handles should not be taken to imply that the *consequences* of wiggling the handles cannot be simultaneous—otherwise we exclude simultaneous causation by fiat.

These considerations play out in two ways. First, and more directly, they suggest that the consequences of lawlike future constraints would be nothing like a future-directed analogue of what we are trying to explain with respect to the past: the sheer apparent absurdity, at least in ordinary cases, of acting so as to influence the past. As we have seen, remote constraints provide little inclination to fatalism, and while immediate constraints would certainly restrict our choices, they would also give us new options.

Second, the agrument suggests that *microscopic* effects on the distant past—which AKL allow to be a consequence of their view—cannot be prevented from being magnified into less microscopic effects on the less distant past, and the remote present, by means of a zig-zag.[16] The engine of the second stage of this process—the 'zag' by means of which the influence of a present action returns from the distant past—will be the very process of amplification of small differences which is central to the account's own proposal concerning macroscopic branching. Suppose it is true (as the AKL account allows) that, had I lifted my little finger a moment ago, there would have been differences in the positions of a number of atoms, billions of years in

the past. What changes might the movement of those ancient atoms have wrought, over such a vast period of time? Not changes enough to dispose of me and my little finger, certainly, for I am here, now, by stipulation, in the history in question. But there is no such protection for the rest of my familiar universe, anywhere within the future light cone of those ancient microscopic changes.

3.3 A General Objection to the Third Arrow Strategy?

We conclude that the AKL approach does not yield a satisfactory explanation of the asymmetry of deliberation. Moreover, the argument just outlined suggests a powerful objection to *any* attempt to ground the time-asymmetry of causation on the kind of macroscopic statistical asymmetries we find in our world. As already noted, it seems highly plausible that these asymmetries have their origin in PH. But we have just argued that since FH would not make it absurd to deliberate for future ends, PH cannot explain why we do not deliberate for past ends. So *any* account of the causal arrow that seeks to reduce the time-asymmetry of causation to the kind of asymmetries that derive from PH seems destined to be similarly powerless to explain the time-asymmetry of deliberation—destined, in other words, to share the failings of Hume's proposal in this respect.

This brings us back to a question we deferred in sect. 2. Insect. 1, generalizing from this objection to Hume's view, we formulated the Practical Relevance Constraint: an account of the time-asymmetry of causation should be expected to explain the time-asymmetry of deliberation. In sect. 2, we observed that it is not *obvious* why we should care about counterfactuals in deliberation in the first place, and hence how Lewis's account might deal with PRC (even if it succeeds in accounting for the time-asymmetry of counterfactual dependence). We now return to that issue.

4. WHY CARE ABOUT COUNTERFACTUALS?

Can Lewis's account meet PRC? Alternatively, can it maintain that PRC is an optional matter for a satisfactory account of causation? Interestingly, these issues have been on the table for many years, in a different guise. There is a long-standing debate between two rival accounts of rational decision, *causal* decision theory (CDT) and *evidential* decision theory (EDT); and a much-discussed class of cases, known generically as *Newcomb problems*, in which the two theories seem to give different recommendations.

The original Newcomb problem (see Nozick 1969) goes like this: we are presented with two boxes, one transparent and one opaque. The transparent box contains $1,000, and we are told that the opaque box may contain either $1,000,000 or nothing. We are offered the choice of taking only the opaque box, or taking both boxes. It seems obvious that we should take both boxes, for that way we are $1,000 better off, *whatever* the opaque box contains. However, we are also informed that the choice of what to put in the opaque box is made by an infallible (or almost infallible) predictor, who places $1,000,000 in the opaque box if and only if he predicts that we will take *only* that box. This information seems to imply that if we take just the opaque

box it is very likely to contain $1,000,000; whereas if we take two boxes, the opaque box probably contains nothing. Doesn't it now make sense to take just one box? Isn't a high probability of $1,000,000 much better than a high probability of $1,000? No, says the rival decision principle, for our choice won't *affect* what is in the opaque box—and whatever it is, we're $1,000 ahead if we take both.

Thus 'one-boxers' argue that we should be guided solely by *evidential* considerations (i.e. by EDT), while 'two-boxers' claim that rationality dictates that we consider *causal* or *counterfactual* considerations (as required by CDT). (Lewis himself was a prominent two-boxer.) The connection with our present concerns is that the issue raised by PRC is a more general form of the issue that divides one-boxers and two-boxers. After all, Newcomb problems are precisely problems in which, according to one-boxers, it is appropriate to act for the sake of an end that one does not *cause*—for example, to raise the evidential probability that the predictor has placed $1,000,000 in the opaque box. The two-boxer's task is to explain why such a decision policy is irrational. And the danger, from the two-boxer's point of view, is that whatever he says about the meaning of cause and effect, the one-boxer is going to respond: 'But if that's what these terms mean, then what's wrong with acting for a end which is *not* an effect of one's action?' This is exactly the challenge that PRC raised to Hume's view.

Thus for a view such as Lewis's, a successful response to PRC and a successful defence of two-boxing would amount to much the same thing. What does the history of these debates tell us about the prospects for such a defence? It reveals a widespread acceptance, even on the part of two-boxers themselves, that there is no such argument to be found. Lewis himself remarks that the debate 'is hopelessly deadlocked' (Lewis 1981*a*: 5). Elsewhere, he puts it like this (Lewis 1981*b*: 378): '[I]t's a standoff. We [two-boxers] may consistently go on thinking that it proves nothing that the one-boxers are richly pre-rewarded and we are not. But [one-boxers] may consistently go on thinking otherwise.'[17]

These remarks support the following assessment of the status of PRC for Lewis's view of causation (and, apparently, for any other view with a similar investment in the issue between CDT and EDT). On the one hand, such views cannot set aside PRC, for they are heavily committed to the relevance of causation to rational deliberation. On the other hand, they have nothing better to offer than a blunt appeal to intuition, in response to the challenge posed by PRC (or, what comes to the same thing, by the one-boxer's challenge to CDT).

In the present context, our interest is in the asymmetry of causation and deliberation. Our reason for mentioning Newcomb problems was that they illustrate so strikingly the gap between proposing an explanation of the causal asymmetry and providing an explanation of the asymmetry of deliberation. One-boxers personify the challenge of PRC, by defending a conception of deliberation that doesn't keep step with causation, at least as ordinarily construed.[18]

But Newcomb problems hold a second message for our present concerns. Why are real-life Newcomb problems comparatively rare, and arcane? Largely, apparently, because even *evidential* deliberation displays a marked temporal asymmetry. If this were not so, after all, then the many cases there would then be of evidential deliberation about past ends would themselves be Newcomb problems. The realization that it is so raises an interesting puzzle, and an inviting prospect. The puzzle is how to characterize and explain this purely evidential asymmetry of deliberation—an asymmetry of an epistemic and 'pre-causal' kind, presumably.

The prospect is that once we have succeeded in doing so, we might have the basis for an understanding of causation itself—an understanding that, by incorporating some of the structure of the epistemic perspective, would gain the means to explain the two things that have so far proved illusive: the temporal orientation of causation, and its relevance to deliberation.

5. THE TIME-ASYMMETRY OF MATERIAL DELIBERATION

Consider a typical case in which we believe that *if* we perform an action *A* (which we take to be within our power to perform or not to perform), an outcome *O* will occur; and in which we don't have reason to think that *O* will occur, independently of whether we perform *A*. Interpreted in material terms, what we believe is simply that the disjunction ¬*A* ⋁ *O* is true. Moreover, we believe it *inferentially*, as we might say—that is, not simply in virtue of already believing one or other disjunct to be true.[19]

Let's call disjunctions of this form—disjunctions held true on inferential grounds, such that the truth of one disjunct is held to be a matter of future choice—*action-linked inferential disjunctions* (ALIDs for short). Here's a striking fact about ALIDs. They are common in cases in which the outcome disjunct (*O*, in our example) concerns a time *after* that of the action disjunct; rare, or perhaps even unknown, in cases in which it concerns a time *before* that of the action disjunct. Call this the *temporal asymmetry of disjunctive deliberation* (TADD).

In the present context, the relevance of TADD is that it reveals a temporal asymmetry that on the one hand is closely linked to deliberation, and on the other seems entirely epistemic in nature—a temporal asymmetry in our typical pattern of disjunctive beliefs about the *actual* world, in cases in which one disjunct concerns one of our own future actions. As we noted, this implies that an account of the causal asymmetry in terms of the counterfactual asymmetry will be blind to at least one significant aspect of the deliberative asymmetry. More intriguingly, it also holds out the prospect that if we could explain TADD then we could also explain everything that needs to be explained about the asymmetries of counterfactuals and causation, if these could be grounded on epistemic or disjunctive deliberation.[20]

Against the latter proposal, it may be objected that there are familiar reasons for distinguishing epistemic from counterfactual deliberation, and for preferring the latter when the two come apart. After all, the former corresponds to one-box reasoning, the latter to two-box reasoning. The epistemicist argues that he knows that he'll have $1,000,000 if and only if he takes one box; the counterfactualist that if he were to take both boxes, he would be $1,000 richer than if he were to take just one box.[21] But our point is that the present context suggests a novel argument on behalf of one-boxing in these debates. In the present context, even a two-boxer needs to explain TADD—and the two-boxer, of all people, must insist that this is a different matter from explaining the analogous asymmetry of counterfactual reasoning. So TADD is a two-boxer's problem, too. Two-boxers have two temporal asymmetries to explain, in effect, whereas a one-boxer has the prospect of an argument that TADD is the *only* asymmetry we need to account for the asymmetry of deliberation.

5.1 What About Cartwright?

It may seem that this prospect is a poor one, in that it collides head-on with the message of a famous paper by Nancy Cartwright (1979). Cartwright argues that causal notions are needed to ground an important distinction between effective and ineffective decision strategies. She describes cases in which evidential and causal deliberation (i.e. EDT and CDT) seem to come apart, and in which it is simply *obvious* that rationality goes with the latter. How, then, could the former kind of deliberation possibly ground the latter?

Our answer is in two parts. First, we note that as subsequent discussion of the kind of decision problems introduced by Cartwright's paper has shown, clear cases are hard to find. Cartwright's examples include so-called 'medical' Newcomb problems, such as one based on the hypothesis that there is a 'smoking gene' that predisposes both to smoking and to cancer. In this case, Cartwright's argument is that a decision to smoke would be evidence that one has the gene, and hence that one has a higher chance of cancer; but that it would clearly be irrational to refrain from smoking on those grounds, if it is what one would otherwise prefer to do.

In such cases, however, it turns out to be far from clear that a rational agent who believes the smoking gene hypothesis should take her own decision to smoke to be evidence that she herself has the gene. Arguably, her knowledge of distinctive features of her own case renders invalid an application of the relevant statistical generalizations (e.g. that most smokers have the gene) to her own decision.[22] If so, then the obvious irrationality of not smoking in this situation rests on faulty evidential reasoning, not on any difference between the recommendations of EDT and CDT. Give EDT the right probabilities, and it, too, recommends that one should smoke.

The remaining cases are both more extreme and far less realistic. For example, they ask us to imagine an agent who has statistical data even about the choices of agents 'just like herself', who have faced exactly her present choice. These cases are much more like the classic Newcomb problem. As well as being highly unrealistic, they share with the classic case the ability to confront us with a deep conflict between seemingly rational intuitions. Hence they are far from clear counterexamples to the approach we are now exploring.

Second, we want to stress that Cartwright's examples *cannot* be clear cases, at least on reflection, if the notion of an effective strategy is to be tied to that of causal or counterfactual reasoning. For in that case, as we have urged, PRC demands an answer. If 'effectiveness' means...—here plug in your favourite causal or counter-factual story—then why should we care about it? Why not be satisfied with an 'ineffective' but probability-raising strategy?

Cartwright is thus in much the same boat as Lewis. On the one hand, she is heavily invested in the link between causation and rational deliberation, and so cannot afford to set PRC aside as irrelevant to an account of causal asymmetry. On the other hand, as is revealed both by the inconclusiveness of arguments for two-boxing in the classic Newcomb problem and by the inability of appeals to PH and the thermodynamic asymmetry to account for the stark asymmetry of deliberation, she has very little prospect of a satisfactory response to PRC.

Far from providing a major obstacle to the suggestion that epistemic deliberation be made the basis of everything else, Cartwright's argument thus provides another illustration of how much is to be gained, if the epistemic approach can be made to work. To do so, however, it needs to find an explanation of the temporal asymmetry of material deliberation (without

appealing to a primitive causal asymmetry, of course). We now turn to this project.

5.2 Explaining TADD

How are we to explain the asymmetry of disjunctive deliberation? A good first question is whether the deliberative aspect—that is, the fact that concerns disjunctions one disjunct of which we take under our control—is likely to play any crucial role. Or does the asymmetry persist if we move to a slightly larger class of disjunctions, without this restriction?

It is easy to see that the asymmetry does not hold if we impose no restriction at all on the form of the disjuncts. Trivially, any disjunction of the form $X \lor Y$ in which one disjunct concerns matters later in time than the other disjunct is equally a disjunction of which the temporal inverse holds. Following the lead of the AKL approach, however, we might suspect that the asymmetry re-emerges when one disjunct concerns a small, local matter, and the other something larger. In this case, too, a material version of AKL might suggest, disjunctions held true on inferential grounds are always such that the 'small local' disjunct concerns a matter earlier in time than the other disjunct.

This simply isn't true, however. After all, consider disjunctions relating forensic evidence (say) to the past states of affairs for which it is evidence. Small differences in the evidence may be indicative of very different histories at earlier times—that's *why* we pay such close attention to forensic details, of course. Thus it may be true, for example, either (S) that a silver medallion just found in the sand does *not* bear the tiny inscription 'CG 1753', or (T) that this beach is the long-lost last resting-place of Captain Greybeard (the oldest sea-dog of his day)—and we may believe $S \lor T$ on inferential grounds—despite the fact that T concerns a matter much earlier than S.

What isn't normally the case, of course, is that we hold true such a disjunction on inferential grounds, *and* believe that the truth of the later disjunct is under our control. (We might believe that whether the medallion bears the inscription 'CG 1753' is under our control, in the sense that we could easily have the inscription added or removed, but in this case we don't hold the disjunct itself true, at least not on inferential grounds.) So the restriction to deliberative cases is crucial to TADD—which raises the question: Is there something temporally asymmetric about agency, about our own deliberative standpoint, that might account for the fact that it seems to introduce an asymmetry in these disjunctive cases that wasn't present without it?

5.3 The Asymmetry of Agency

We have just observed that we can't use evidence as a 'causal handle' to influence the earlier states of affairs for which it provides evidence. This suggests that the distinguishing feature of causal handles isn't a temporal-direction-neutral fact about the correlation of small local differences with big remote differences. On the contrary, it seems to lie in the simple fact that we can only wiggle handles that lie in the immediate *future,* with respect to our own deliberations on the matter. If this is right, then the source of the temporal asymmetry of TADD is our own asymmetric perspective as agents—the fact that we are always

contemplating actions in the near *future*, with respect to the time of deliberation—not some independent fact about the structure of reality.

Looking at this from the point of view of the matters we contemplate bringing about in deliberation, this asymmetry plays out in a marked temporal asymmetry in associated states of affairs, in the immediate temporal vicinity of the matters in question. To think of the matters we bring about *as* products of deliberation is to think of them as *having a particular history*—as being immediately *preceded* by our own deliberation, in effect. This makes a huge difference to their evidential significance in that direction, of course, as our last example illustrates: the evidential bearing on past states of affairs of the presence of an inscription on an old medallion is highly sensitive to whether we have just chosen to put it there.

In other words, the very presence of deliberation ensures that the events contemplated in deliberation are *not* typical as regards their associations in the past—for in the past lies the deliberation itself. And yet there is no such restriction in the future. No wonder, then, that that inference *from* the fact of the occurrence of such an event should work so differently in the two temporal directions.

The crucial difference here, compared to the AKL approach, is that we have shifted from considering the evidential consequences of small, local changes *in general*—wiggles of 'causal handles', or local changes produced by agents with *arbitrary* temporal orientation—to thinking of those such changes that are the products of deliberation by agents with *our* temporal orientation: agents for whom actions *follow* deliberation, in the usual time sense. In an account of this kind, then, the asymmetry is being supplied by the asymmetry of our own particular deliberative standpoint, rather than by an objective asymmetry such as PH.[23] It is thus analogous to cases such as those we mentioned at the end of sect. 1: the distinctions between near and far, or local and foreigner, or left and right.

So far, we are talking about TADD, and hence about cases in which changes are thought of as possible actions. For the moment, the claim is simply that the temporal asymmetry of the deliberative standpoint itself does a good job of accounting for TADD. If we are to make the further claim that the asymmetry of the deliberative standpoint underlies that of counterfactuals and causation in general, it needs to be explained how we are to make the step from this restricted case to the general case—if the asymmetry of the deliberative standpoint is to do the work in the general case, it will need to be argued that when we assess counterfactuals, we think of the antecedents *as* potential actions, with the asymmetry intact. We'll return to this issue in a moment. First, before we leave the relative simplicity of the disjunctive case, it is worth asking whether TADD itself is a strict temporal asymmetry, or whether the account allows for backward-directed—'retroactive'—disjunctive deliberation.

5.4 Retroactive Disjunctive Deliberation?

Retroactive disjunctive deliberation ('RetroDD') seems to exist in two varieties. The first is illustrated by our modified Death in Damascus example, from sect. 3.2. In this case, we believe a disjunction of the form:

$$\text{(We will stay in Aleppo)} \bigvee \text{(Death is already on his way to Damascus)} \quad (1)$$

We believe it on inferential grounds, and we take the first disjunct to be one that we can decide to make true or false, as we wish. So the case meets the criteria for disjunctive deliberation, despite the fact that the second disjunct concerns a time in the past, relative to that of the deliberation. Let's call the pattern exemplified here *zigzag* RetroDD—it turns on the fact that something we can choose to make the case in the future is suitably correlated with a state of affairs in the past (even in the circumstances in which we take ourselves to have the choice).

The second kind of RetroDD—in some sense, a limiting case of the first—is where we take *our choice itself* to be correlated with an earlier event. This is the case associated with medical Newcomb problems, such as the smoking gene example from sect. 5.1. Consider the extreme version, in which the hypothesis is that all and only those who have the gene become smokers. For someone who believed both that this correlation holds, and that he nevertheless had a choice as to whether to smoke, the following disjunction, too, would meet the relevant criteria:

$$\text{(I will not smoke)} \bigvee \text{(I have the cancer gene)} \quad (2)$$

Of course, it is hard to imagine why someone should combine both the required beliefs. Prima facie, they seem to be in tension. (Perhaps the original Newcomb problem does as good a job as can be done of presenting a case in which it seems reasonable that we might believe both.) But for the moment, what matters is simply that for someone who did combine them, the result would be an example of RetroDD—we might call it *direct* RetroDD.

We emphasize again that this discussion has been confined to the epistemic case. At this stage, counterfactuals and causal reasoning are simply not in the picture. But the fact that disjunctive deliberation allows, at the margins, for these retroactive cases implies that if epistemic deliberation can be made the foundation for counterfactual deliberation, then it, too, stands to inherit the same temporal character: overwhelmingly 'past-to-future', though with loopholes for exceptional cases. And as we noted at the beginning, this seems to be precisely what we want of an account of the temporal asymmetry of causation.

6. THE ATTRACTIONS OF SUBJECTIVISM

At the end of sect. 1 we observed that if we reject two extreme views—Hume's conventionalism and hyperrealism—we seem to be left with two options for explaining the nature and temporal orientation of the causal arrow. The first, the third arrow approach, looks for some objective physical asymmetry to ground the causal asymmetry. We argued that the only apparent candidate, some sort of de facto statistical asymmetry linked to the thermodynamic asymmetry, seems unpromising. For one thing, it reduces to something very much like Hume's view in the case of microscopic and substatistical systems, where the causal asymmetry becomes nothing more than a conventional label, applied to mark alignment

with a macroscopic statistical asymmetry. For another, its link to deliberation is at best obscure. In particular, the statistical asymmetry does a poor job of explaining why we don't (typically) deliberate with respect to past ends.

The second option, we noted, is to agree with Hume that there is no intrinsic asymmetry of causation, but to look for some better story than Hume's own about why our causal notions show such a strong and temporally asymmetric asymmetry. In sects. 4–5 we have been investigating the credentials of one obvious candidate for the beginnings of such a story, namely, our own perspective as agents and deliberators. We have discovered that if we think of deliberation, initially, in epistemic, evidential, or 'pre-causal' terms, it nevertheless exhibits a strong temporal asymmetry: an asymmetry explicable, apparently, in terms of our own asymmetric temporal orientation, as 'players' in the dynamical environments in which we live; and an asymmetry that allows, at the margins, for the epistemic analogue of retrocausality.

This is a very striking result. If it could be elaborated into a plausible explanation (or better, *genealogy*) of our ordinary causal concepts, and of associated matters, such as counterfactual reasoning, it would tick all the hard boxes, apparently. It wears its link with deliberation on its face, so there are no problems with PRC. It has good physicalist credentials so long as the notion of agency itself does: in other words, so long as biology and physics can account for the existence of creatures like us; and it links to the thermodynamic asymmetry so long as that explanation does so. It gets the character of the temporal asymmetry just about right: predominantly though perhaps not universally past-to-future, in our time sense (and plausibly linked to de facto physical asymmetries, for the reason just mentioned). It gets the scope of the causal asymmetry just about right, too, in the sense that so long as our deliberative perspective is blind to the micro–macro distinction, then so is the causal asymmetry. And it makes it immediately obvious, in a way that Hume's own conventionalism does not, why we have an interest in marking (what we come to call) the cause–effect distinction: we thereby mark something of first importance, from an agent's perspective.

Despite these advantages, many philosophers feel that this approach to the causal asymmetry gives away too much: it renders the causal asymmetry insufficiently objective. It is worth noting, however, that there is one sense in which this battle has already been lost. The main rival, the statistical view, has already conceded that there is no instrinsic asymmetry at a fundamental level. Critics thus do better to focus their attention on the challenges of the project of turning the subjectivist's proposed raw materials into a plausible genealogy for our causal concepts and cognitive machinery.

We cannot explore the prospects for that project here, but we close with a suggestion about how to think of the 'subjectivism' of this view, and with two notes about how it might tie in in interesting ways with aspects of the theory of causation normally thought of in other ways.

6.1 A Subjectivist's Guide to Objective Causation?

The project is to ground the asymmetry and practical relevance of causation on that of deliberation, epistemically construed. This idea seems strikingly analogous to a viewpoint long familiar in the case of probability. In that case, probabilistic 'subjectivists' are united by

the thought that a proper account of probability needs to begin on the practical and epistemic side—that is, with *credence,* defined in terms of its role in decision under epistemic uncertainty. Not all subjectivists think of this as incompatible with recognizing more objective notions of probability as well, but their common motto is that if an account of probability doesn't build the link with decision in at the beginning, it will never be able to recover it later —never be able to justify the link between objective probability and credence that Lewis calls the Principal Principle.[24]

We suggest that the lesson of PRC be viewed in the same light, and be called 'subjectivist' for the same reason. Indeed, PRC itself seems to play a role analogous to that of the Principal Principle. And subjectivism here consists in reading its implications in a similar way: unless an account of causation starts with deliberation, epistemically construed, it is not going to be able to explain why causation matters to deliberation in the way that it does. As in the case of probability, this starting point leaves room for a range of possible views, at the more objectivist end of which might be a causal analogue of Lewis's view of chance. But what these views will have in common will be a recognition that for causation, as for probability, the practical, epistemic perspective is importantly prior to the metaphysical perspective.

6.2 Folk Physics and the Fixity of the Past

We noted earlier that Lewis observes that one might treat the asymmetry of counterfactual dependence as the product of a convention—a stipulation that when we assess counterfactuals, we 'hold the past fixed'. He rejects this option for much the same reasons that many philosophers reject Hume's conventionalism, for example, that it puts the asymmetry in by hand, and rules out backward dependency by fiat. But the subjectivist view gives new interest to the idea that counter-factual reasoning might be governed by such a convention. If the relevant species of counterfactual reasoning develops from the kind of hypothetical reasoning needed in epistemic deliberation, the principle that one should hold the past fixed provides a simple codification of the asymmetry of the deliberator's perspective—a codification that won't lead to problems, apparently, so long as the environment does not supply the kind of rare opportunities that might favour retroactive deliberation.

Hence it is tempting to suggest that the fixity of the past has the status of a useful piece of folk physics, deeply ingrained as our ancestors developed the cognitive framework that supports deliberation. With this hypothesis in place, subjectivists are free to help themselves to an asymmetry of counterfactual dependence, grounded on the (now explicable) convention that Lewis rejects; and hence, if they wish, to the resources of a counterfactual account of causal reasoning. They are also free to discuss possible modifications in the folk physics, for example, to accommodate retrocausality in the kinds of cases to which Lewis himself calls our attention (see n. 1).[25]

6.3 Interventionism

Much recent work has focused on links between causation and what has come to be called

intervention. Roughly, an intervention is a 'surgical' input into a system of correlated variables that sets the value of a particular variable, breaking the normal links between it and its causal 'parent'. As Woodward (2001) puts it: '[T]he intervention disrupts completely the relationship between [a variable X] and its parents so that the value of [X] is determined entirely by the intervention. Furthermore, the intervention is surgical in the sense that no other causal relationships in the system are changed.' The basic proposal is, then, that the effects of X are the dependencies that survive when the value of X is fixed by an intervention of this kind. As Woodward goes on to note, this may be seen as a formalization of the central idea of manipulability approaches to causation, such as that of Menzies and Price (1993): 'In this way, we may capture Menzies' and Price's idea that X causes Y if and only if the correlation between X and Y would persist under the right sort of manipulation of X.'

It seems clear that this connection will be of great importance to any attempt to develop a subjectivist approach to the causal asymmetry. Ideally, the subjectivist will want to step into the interventionists' shoes—all the more so, now that Pearl, Woodward, and others have shown us how far those shoes may take us![26] On the face of it, the shoes seem to fit extremely well. The defining feature of an intervention is that it breaks 'upsteam' links, in a very similar manner, apparently, to the way in which we have seen that the mere presence of the deliberating agent breaks links to the past, in the case of disjunctive deliberation. In a sense, the main issue is who owns the shoes in the first place. Does deliberation need to be explained as a species of intervention, in other words, or is deliberation the primary notion?

Subjectivists put their money on the latter option, and we close by noting two sorts of argument they may offer, drawing on the conclusions of our earlier discussion. The first claims that only our contingent temporal asymmetry as agents can account for the fact that the class of interventions relevant to ordinary causal judgements are interventions 'from the past', not 'from the future'. (As before, subjectivists claim that their view explains an asymmetry that other views must treat as primitive, or simply leave unexplained.)

The second argument appeals to PRC, and turns the tables on popular objections to subjectivism in an interesting way. It is often objected that manipulability theories of causation will be circular, because manipulation is a causal notion. But we have now seen that deliberation can be characterized in a non-causal, epistemic fashion. As long as deliberation is construed in epistemic terms, in other words, it is simply not true that the manipulability theory relies on a causal notion at this point.[27] Whereas if intervention is the basic notion, then not only does it rely for its characterization on causal notions, rendering circular any *analysis* of causation in interventionist terms,[28] but this also leaves it vulnerable to the challenge of PRC. What is it about *that* causal notion—whatever it is—that renders it relevant to deliberation?[29]

7. SUMMARY

There is a considerable consensus that there is no fundamental, intrinsic asymmetry of causation. To that extent, Hume and Russell seem to have been right: there is no asymmetric causation in Sellars's 'scientific image', at least at its most basic level. Concerning the 'manifest image'—the explanation of the asymmetry and temporal orientation of ordinary

concepts and judgements about causation, and of related matters, such as deliberation and counterfactual reasoning—the most promising strategy seems to be to begin with the de facto asymmetry of human deliberation, characterized in epistemic terms, and to build out from there. More than any rival, this subjectivist approach promises to demystify the asymmetry, temporal orientation, and deliberative relevance of our causal judgements.

In a recent survey article about causation, much concerned with the issue of temporal asymmetry, Hartry Field (2003: 443) remarks: '[W]e have a problem to solve: the problem of reconciling Cartwright's points about the need of causation in a theory of effective strategy with Russell's points about the limited role of causation in physics. This is probably the central problem in the metaphysics of causation.' We have suggested, in effect, that the best option is to move the problem from metaphysics to pragmatics. So long as we see the problem as one of explaining the practical relevance of causal notions, in the lives of creatures in our situation, there is some prospect of reconciliation.

FURTHER READING

Reichenbach (1956) developed the first third arrow account, grounded in a probabilistic theory of causation. Horwich (1987) and Hausman (1998) are more recent theories developed along broadly similar lines. Lewis (1979) first proposed the counterfactual overdetermination account, to which Elga (2000) provides an important objection. Field (2003) is a useful survey covering all these accounts. Recent examinations of causal asymmetry in the context of fundamental physical theories are Price (1996), Albert (2001), and Frisch (2005*b*). Finally, Price and Corry (2007) is a collection of recent papers on causation and physics, many of which address the issues at hand.

REFERENCES

ALBERT, DAVID Z. (2001). *Time and Chance*. Cambridge, Mass.: Harvard University Press.

ARNTZENIUS, FRANK (1990). 'Physics and Common Causes', *Synthese* 82/1: 77–96; http://dx.doi.org/10.1007/BF00413670>, accessed 11 March 2009.

CALLENDER, CRAIG (2004). 'Measures, Explanations and the Past: Should "Special" Initial Conditions be Explained?', *British Journal for the Philosophy of Science* 55/2: 195–217; http://dx.doi.org/10.1093/bjps/55.2.195>, accessed 11 March 2009.

CAMPBELL, RICHMOND, and SOWDEN, LANNING (1985). *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem* . Vancouver: University of British Columbia Press.

CARTWRIGHT, NANCY (1979). 'Causal Laws and Effective Strategies', *Noûs* 13/4: 419–37.

COLLINS, JOHN, HALL, NED, and PAUL, L. A. (2004). *Counterfactuals and Causation*. Cambridge, Mass.: MIT.

DUMMETT, MICHAEL (1954). 'Can an Effect Precede Its Cause?', *Proceedings of the Aristotelian Society Supplement* 28/3: 27–44.

—— (1964). 'Bringing About the Past', *Philosophical Review* 73/3: 338–59.

EELLS, ELLERY (1981). 'Causality, Utility, and Decision', *Synthese* 48/2: 295–329; repr. in Eells 1982; http://dx.doi.org/10.1007/BF01063891>, accessed 11 March 2009.

—— (1982). *Rational Decision and Causality*. Cambridge: Cambridge University Press.

ELGA, ADAM (2000). 'Statistical Mechanics and the Asymmetry of Counterfactual Dependence', *Philosophy of Science* 68/3: S313–S324; http://dx.doi.org/10.1086/392918>, accessed 11 March 2009.

FEYNMAN, RICHARD (1965). *The Character of Physical Law*. Cambridge, Mass.: MIT.

FIELD, HARTRY (2003). 'Causation in a Physical World', in Michael J. Loux and Dean W. Zimmerman (eds.), *The Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 435–60; http://philosophy.fas.nyu.edu/docs/IO/1158/Cause.pdf>, accessed 11 March 2009.

FRISCH, MATHIAS (2005*a*). 'Counterfactuals and the Past Hypothesis', *Philosophy of Science* 72/5: 739–50; http://dx.doi.org/10.1086/508111>, accessed 11 March 2009.

—— (2005*b*). *Inconsistency, Asymmetry, and Non-locality: A Philosophical Investigation of Classical Electrodynamics*. Oxford: Oxford University Press; http://dx.doi.org/10.1093/0195172159.001.0001>, accessed 11 March 2009.

—— (2007). 'Causation, Counterfactuals, and the Past-Hypothesis', in Price and Corry (2007).

—— (forthcoming) 'Does a Low-Entropy Constraint Prevent Us from Influencing the Past?', in Gerhard Ernst and Andreas Hüttemann (eds.), *Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics*. Cambridge: Cambridge University Press; http://philsci-archive.pitt.edu/archive/00003390/>, accessed 11 March 2009.

GIBBARD, ALAN, and HARPER, WILLIAM (1978). 'Counterfactuals and Two Kinds of Expected Utility', in Clifford Hooker, James Leach, and Edward McLennen (eds.), *Foundations and Applications of Decision Theory Foundations and Applications of Decision Theory*. Dordrecht: D. Reidel, 125–62.

HAUSMAN, DANIEL M. (1998). *Causal Asymmetries*. Cambridge: Cambridge University Press.

HAWKING, STEPHEN W. (1994). 'The No Boundary Condition and the Arrow of Time', in Jonathan J. Halliwell, Juan Pérez-Mercader, and Wojciech Hubert Zurek (eds.), *Physical Origins of Time Asymmetry*. Cambridge: Cambridge University Press, 346–57.

HORGAN, TERENCE (1981). 'Counterfactuals and Newcomb's Problem', *Journal of Philosophy* 78/6: 331–56.

HORWICH, PAUL (1987). *Asymmetries in Time: Problems in the Philosophy of Science*. Cambridge Mass.: MIT.

KUTACH, DOUGLAS (2001). 'Entropy And Counterfactual Asymmetry'. Ph.D. thesis. New Brunswick, NJ: Rutgers University; http://www.brown.edu/Departments/Philosophy/Douglas_Kutach/Kutach-Dissertation.pdf>, accessed 11 March 2009.

—— (2002). 'The Entropy Theory of Counterfactuals', *Philosophy of Science* 69/1: 82–104; http://dx.doi.org/10.1086/338942>, accessed 11 March 2009.

—— (2007). 'The Physical Foundations of Causation', in Price and Corry (2007); http://www.brown.edu/Departments/Philosophy/Douglas_Kutach/Kutach-PhysicalFoundationsofCausation.pdf>, accessed 11 March 2009.

—— (forthcoming). 'The Asymmetry of Causal influence', in Craig Callender (ed.), *Oxford Handbook of Time*. Oxford: Oxford University Press.

LANGE, MARC (2006). *Philosophy of Science: An Anthology*. Malden, Mass.: Blackwell.

LEWIS, DAVID (1973). 'Causation', *Journal of Philosophy* 70/17: 556–67; repr. in Lewis 1986*a*: 159–71; http://dx.doi.org/10.2307/2025310>, accessed 11 March 2009.

—— (1979). 'Counterfactual Dependence and Time's Arrow', *Noûs* 13/4: 455–76; repr. in Lewis 1986*a*: 32–52; page references are to the latter version. http://dx.doi.org/10.2307/2215339>, accessed 11 March 2009.

—— (1980). 'A Subjectivist's Guide to Objective Chance', in Richard C. Jeffrey (ed.), *Studies in Inductive Logic and Probability*. Berkeley: University of California Press, ii. 263–93; repr. with postscripts in Lewis 1986*a*: 83–132; http://dx.doi.org/10.1093/0195036468.003.0004>, accessed 11 March 2009.

—— (1981*a*) 'Causal Decision Theory', *Australasian Journal of Philosophy* 59/1: 5–30; repr. in Lewis 1986*a*: 305–39; page references are to the latter version.

—— (1981*b*). '"Why Ain'cha Rich?" ', *Noûs* 15/3: 377–80.

—— (1986*a*). *Philosophical Papers II*. New York: Oxford University Press.

—— 1986*b*. 'Postscripts to "Counterfactual Dependence and Time's Arrow" ', in Lewis 1986*a*: 52–66.

—— (2000). 'Causation as Influence', *Journal of Philosophy* 97/4: 182–197; repr. in Collins, Hall, and Paul 2004: 75–106 and Lange 2006: 466–87. http://dx.doi.org/10.2307/2678389>, accessed 11 March 2009.

LOEWER, BARRY (2007). 'Counterfactuals and the Second Law', in Price and Corry 2007: 293–326.

MENZIES, PETER, and PRICE, HUW (1993). 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science* 44/2: 187–203.

NOZICK, ROBERT (1969). 'Newcomb's Problem and Two Principles of Choice', in Nicholas Rescher (ed.), *Essays in Honor of Carl G. Hempel*. Dordrecht: Reidel, 114–46; repr. in Nozick 1997: 45–74 and Campbell and Sowden 1985: 107–33.

—— (1997). *Socratic Puzzles*. Cambridge, Mass.: Harvard University Press.

PEARL, JUDEA (2000). *Causality*. Cambridge: Cambridge University Press.

PRICE, HUW (1986). 'Against Causal Decision Theory', *Synthese* 67/2: 195–212; http://dx.doi.org/10.1007/BF00540068>, accessed 11 March 2009.

—— (1991). 'Agency and Probabilistic Causality', *British Journal for the Philosophy of Science* 42/2: 157–76; http://dx.doi.org/10.1093/bjps/42.2.157>, accessed 11 March 2009.

—— (1996). *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*. Oxford: Oxford University Press.

—— and Corry, Richard (2007). *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press.

RAMSEY, FRANK PLUMPTON ([1929] 1931). 'General Propositions and Causality', in Richard B. Braithwaite (ed.), *The Foundations of Mathematics and other Logical Essays*, London: Kegan Paul, Trench, Trübner, 237–55; repr. in Ramsey (1978: 133–51); page references are to the latter edition. http://www.dspace.cam.ac.uk/handle/1810/194722>, accessed 11 March 2009.

—— (1978). *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*, ed. D. H. Mellor. London: Routledge, Kegan Paul.

REICHENBACH, HANS (1956), *The Direction of Time*. Berkeley: University of California Press.

RUSSELL, BERTRAND (1912–13). 'On the Notion of Cause', *Proceedings of the Aristotelian Society* 13: 1–26.

SPIRTES, PETER, GLYMOUR, CLARK, and SCHEINES, RICHARD (2000). *Causation, Prediction and Search*. 2nd ed. Cambridge, Mass.: MIT; http://cognet.mit.edu/library/books/view?isbn=0262194406>, accessed 11 March 2009.

TOOLEY, MICHAEL (1987). *Causation: A Realist Approach*. Oxford: Clarendon.

—— (1990). 'Causation: Reductionism versus Realism', *Philosophy and Phenomenological Research* 50 suppl: 215–36.

WESLAKE, BRAD (2006). 'Review of *Making Things Happen*', *Australasian Journal of Philosophy* 84/1: 136–140; http://dx.doi.org/10.1080/00048400600571935>, accessed 11 March 2009.

WOODWARD, JAMES (2001). 'Causation and Manipulability', in Edward N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. Stanford: Stanford University; http://plato.stanford.edu/entries/causation-mani/>, accessed 11 March 2009.

—— (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press; http://dx.doi.org/10.1093/0195155270.001.0001>, accessed 11 March 2009.

PART V
THE EPISTEMOLOGY OF CAUSATION

CHAPTER 21
THE PSYCHOLOGY OF CAUSAL PERCEPTION AND REASONING

DAVID DANKS

1. INTRODUCTION

Causal beliefs and reasoning are deeply embedded in many parts of our cognition (Sloman 2005). We are clearly 'causal cognizers', as we easily and automatically (try to) learn the causal structure of the world, use causal knowledge to make decisions and predictions, generate explanations using our beliefs about the causal structure of the world, and use causal knowledge in many other ways. Because causal cognition is so ubiquitous, psychological research into it is itself an enormous topic, and literally hundreds of people have devoted entire careers to the study of it. As such, this chapter will necessarily be woefully incomplete. Each of the sections below (except perhaps sect. 4) could easily be expanded to an entire book, and this chapter must (by necessity) leave unaddressed some areas of psychological research that are plausibly relevant to causal cognition.[1]

Causal cognition can be divided into two rough categories: causal learning (sects. 2–4) and causal reasoning (sect. 5). The former encompasses the processes by which we learn about causal relations in the world at both the type and token levels; the latter refers to the ways in which we use those causal beliefs to make further inferences, decisions, predictions, and so on. The two types of causal cognition are clearly connected to one another, but psychological research on each has proceeded relatively independently from the other. Causal learning itself can be divided into two distinct types: causal perception (sect. 2) and causal inference (sect. 3). Causal perception consists of the relatively automatic, relatively irresistible perception of certain sequences of events as involving causation. For example, if a nearby car alarm goes off when I close my own car door, then I cannot help but perceive my own action as causing the alarm, even though I know that my action was not causally relevant. Causal inference, on the other hand, consists of higher-level causal learning that is based largely on statistical relationships. For example, I learn that one drug is better than another for pain relief by considering the relevant (statistical) history. There are historical and sociological reasons for this split in research on causal learning, but there are also apparent differences in phenomenology, behaviour, and underlying neural bases. The precise connection between causal perception and inference is discussed in more detail in sect. 4. Finally, research on non-human animals (sect. 6) has in recent years helped us to understand better the nature of human causal cognition by revealing ways in which our causal cognition is similar to, and differs from, that of other animals.

2. Causal Perception

Consider looking at a computer screen with a red square in the centre, and a green square moving smoothly towards the centre from the left side. Suppose further that the green square stops when it first 'touches' (i.e. is contiguous with) the red square, and the red square begins moving to the right at the same speed. Nothing about this description, or (seemingly) the visual information, indicates anything about causation; this really is nothing more than a sequence of images on a computer screen. Nonetheless, when presented with a sequence of images such as these, almost everyone will immediately and spontaneously say that the green square *caused* the red square to move.[2] No conscious thought or reasoning seems to be required, and very few observers can avoid believing that the one block caused the movement of the other. This canonical instance of causal perception—referred to as the *launching effect*—was first systematically explored by Albert Michotte, and catalogued in his 1946 book *La Perception de la causalité* (translated in 1963). Michotte conducted over one hundred studies investigating exactly when the launching effect does and does not arise spontaneously in observers. He showed, for example, that the standard launching effect does not arise if there is a spatial or temporal gap between the objects' movements: if the green square stops short of the red one, or the red one's movement occurs (noticeably) after the green square touches it, then one does not experience any perception of causality.

Michotte's findings were originally quite surprising, but methodological concerns about his results have essentially all been answered, and the basic phenomenon of causal perception is now widely accepted (White 1995). Even infants are widely thought to experience causal perceptions (Leslie 1982; 1984; Leslie and Keeble 1987; Oakes 1994; Oakes and Cohen 1990). Post-Michotte research on causal perception has largely aimed to determine both the exact conditions under which causal perception occurs and the constituent processes of causal perception. As an example of the first line of research, Scholl and colleagues have shown that causal perception depends on context, and not just the primary objects (Choi and Scholl 2004; Scholl and Nakayama 2002). Suppose that the green square (in the original example) moves over the top of the red square and stops when it completely covers the red square, and the red square only starts moving once it is completely covered. In these cases, people do not normally experience any causal perception; rather, they usually experience one object passing smoothly over another and changing colour spontaneously in the middle of the motion. If, however, an ordinary launching event occurs somewhere else on the screen at the same time, then the experience of this sequence changes. In this case, it is viewed causally as the green square launching the red. Whether an image sequence is perceived causally can thus change depending on seemingly irrelevant events elsewhere in the visual field, and even infants are affected by these sorts of contextual changes (Newman et al. 2008).

Research on the component processes of causal perception has focused on its development, and its neuronal bases. Developmentally, 6-month-old infants seem to perceive the basic launching effect stimuli in terms of causality, rather than 'simpler' perceptual features such as contiguity or persistence (Leslie 1982; Leslie and Keeble 1987).[3] Oakes and Cohen (1990) found that causal perception arises for more complex stimuli (e.g. unusual trajectories) only in 10-month-olds, but not 6-month-olds (see also Oakes 1994). Infants younger than 6 months

old attend to (i.e. perceive) only some of the spatial and temporal components of a launching event; they do not seem to have 'full' causal perceptions (Cohen and Amsel 1998). Causal perception thus seems to require (separable) perceptions of appropriate spatial and temporal contiguity, and also the ability to ignore extraneous perceptual elements. Causal perception does not arise in one fell swoop, but rather comes together more slowly. In adults, spatial contiguity even ceases to be necessary in some cases; causal perceptions can arise without any spatial contact at all between the on-screen objects (White and Milne 1997; 1999). Similar results emerge from the limited neuroscientific work on causal perception: for example, fMRI data suggest that overlapping, but not identical, brain regions are responsible for the spatial and temporal components of causal perception (Fugelsang et al. 2005). There does not seem to be a single, neuronally distinct 'module' for causal perception (though see below).

Most experiments on causal perception have focused on variants of launching events, but causal perception can arise in other contexts. In their classic experiments, Heider and Simmel (1944) showed that the movement of simple geometric objects is sometimes perceived as *intentional* movement by the objects. For example, suppose the red square (in the original launching event) begins moving before the green square arrives, and the red square moves erratically while the green square follows smoothly behind it. This sequence of images will typically be perceived as the red square 'fleeing' while the green square 'chases'. That is, causal perceptions arise not just for physical causation, but also for social or intentional causation, and are similarly automatic and unprompted in the latter domain. These perceptions of objects as 'intentional agents' whose states can cause behaviour seem to arise as early as 9 months old (Csibra et al. 1999; Gergely et al. 1995). Physical and social causal perceptions do appear to be separable, however, as the former seems to be perceived more strongly than the latter (Schlottmann et al. 2006).

Causal perception has traditionally been viewed as philosophically interesting because it seems to be a (partial) psychological vindication of Kant over Hume: certain judgements of causality seem to be part-and-parcel of perception, rather than something that occurs after 'basic' perception has taken place. Moreover, these causal perceptions can influence other perceptual judgements, such as event timing (Choi and Scholl 2006; Newman et al. 2008), and causal perception does not seem to be susceptible to top-down control or overriding (Blakemore et al. 2001; Fonlupt 2003). Causality seems to be built in to some of our perceptions of the world, rather than always being only inferred from a sequence of images. More generally, as Michotte (1963) himself realized, causal perception seems to be a plausible candidate for a modular process (in the Fodor 1983 sense), as it is fast, automatic, mandatory, and informationally encapsulated. Causal perception (seemingly) depends only on visual input, and not on higher-level cognition; you cannot, for example, choose not to see the classic launching events as causal. It also arises cross-culturally, and causal perceptions are the same even for groups that make quite different causal attributions in social contexts (Morris and Peng 1994).

Several researchers have used the above reasons to argue that causal perception is plausibly a cognitive module (Leslie 1984; 1994; Leslie and Keeble 1987; Scholl and Tremoulet 2000), and perhaps even a neurological module (Blakemore et al. 2001). But although causal perception behaves modularly in processing, there are reasons to doubt that it constitutes a fully Fodorian module. It does not have a classically modular development (Schlottmann

2000), as causal perception requires different cognitive components that develop at different times. Neurally, these components seem to be distributed relatively widely: both temporal lobes, the inferior parietal lobe, and the frontal gyrii (Blakemore et al. 2001; Fugelsang et al. 2005). Behaviourally, there do not seem to be any reported cases of selective loss of causal perception; at the current time no individuals have been found with lesions or other neural damage that resulted in loss of causal perception (without much broader loss of visual perception). There are also significant individual differences in causal perception. Examples include the findings that (1) some individuals fail to have a causal perception for classic launching stimuli (e.g. Beasley 1968); (2) causal perceptions or their absence can change upon repeated exposure of the same stimuli; and (3) experience can affect whether causal perceptions occur. Moreover, these individual variations are largely stable over time (Schlottmann and Anderson 1993), and so suggest that the 'module' at least has important parameters that are set by personal experience. Causal perception has many modular features —automaticity, mandatory triggering, and informational encapsulation—but does not seem to satisfy fully the classical profile of a module.

3. CAUSAL INFERENCE

A different type of causal learning occurs when one is learning that exposure to a particular plant (e.g. poison ivy) causes a rash, or that a new drug has various side effects. In these cases, one often cannot rely on spatio-temporal cues, but rather must attend to differences in occurrence rates in some relevant population. The paradigmatic situation for causal inference is one in which the learner observes a series of situations or cases in which various potential causes do or do not occur, and the presumptive effect does or does not occur (Cheng 1997; Cheng and Novick 1990; Shanks 1995). The learning challenge is then to determine which potential causes are actual causes, and also the strengths (in some sense) of those causes. Numerous variations on this paradigmatic situation are obviously possible; for example, one might have spatio-temporal information (e.g. Buehner and May 2002), or the learner might actively bring about some of the cases (e.g. Sobel and Kushnir 2006). The central challenge remains largely the same, however: use principally statistical information (e.g. something like correlation) to learn causal relations and strengths. This type of causal inference is not directly perceptual, nor does it seem to have the same type of automaticity as causal perception: people rarely learn that a plant causes a rash after only one exposure (though they might suspect that it does so). Causal inference also seems to require higher-order (in some sense) cognition than causal perception. As a result, psychological research on causal inference has proceeded relatively independently of research on causal perception (though see sect. 4 below).

The dominant experimental paradigm in psychological research on causal inference has three principal components. First, the 'cover stories' largely prevent experimental participants from using any substantive prior causal knowledge beyond, for example, temporal order. Second, the relevant variables for causal inference are always obvious in the stimuli, and the variables might even be divided into potential causes and an effect. Third, participants provide their judgements about the causal relations as explicit numeric ratings for each potential

cause, usually on a scale ranging from − 100 ('always prevents') to +100 ('always generates'). There are of course experiments without one or another of these components: for example, the cover story might evoke substantive domain knowledge (e.g. Schulz and Gopnik 2004); the cases might be presented using actual objects (e.g. Gopnik et al. 2004); participant behaviour might generate the cases (e.g. Buehner and May 2003; Steyvers et al. 2003); or participants might respond with graphs rather than numeric ratings (e.g. Steyvers et al. 2003). The principal theoretical task, however, is surprisingly constant over all of these variations: explain the patterns of ratings that are generated by systematic variations in the statistical relationship between the potential causes and effect.

One intuition explored early in the psychological research is that human causal inference might be similar to, or even identical to, the associative learning processes found in non-human animals. Most people are familiar with the notion of classical (Pavlovian) conditioning: repeatedly ring a bell (referred to as the cue or Conditioned Stimulus, CS) just before presenting a dog with food (the outcome or Unconditioned Stimulus, US) and the dog will come to associate bell-ringing with food (and so salivate upon bell-ringing). Instrumental conditioning refers to situations in which the relevant cue is generated through the animal's own action (e.g. the dog presses a lever). Of course, both types of conditioning can lead to quite complex patterns of behaviour, as numerous behaviourist experiments demonstrated (e.g. Skinner's pigeons that famously 'played' ping-pong). Broadly speaking, formal models of these processes, and the full range of conditioning phenomena, are referred to as *associative* models (though in recent years, 'associative' has frequently been used as a pejorative term to refer to any model that an author does not like). The dominant associative model of the past thirty years is the Rescorla–Wagner (1972) model, though many alternatives have been proposed (e.g. Pearce 1987; Van Hamme and Wasserman 1994). All the models represent associative learning as the learning of so-called associative strengths for the different factors, and they share other features: the processes require little memory or computational power; cases are handled sequentially, rather than as a group; and learning proceeds through an error-correction process (i.e. associative strengths are adjusted based on the error between their prediction about whether the outcome will occur and whether it actually does occur). There are also models that share these features, though they have no history in the animal behaviour literature (Catena, Maldonado, and Candido 1998; Danks, Griffiths, and Tenenbaum 2003). One proposal for causal inference is that the causal strengths learned in human causal inference (and reported as ratings in experiments) might actually be associative strengths learned using some associative process (Shanks 1995).

Instead of the case-by-case learning characterized by associative models, one could focus on asymptotic causal inference: what causal relations do people learn after a long enough sequence or summary of cases (i.e. once their beliefs stabilize)? This type of causal inference is closely connected to, but importantly different from, the narrower problem of contingency learning: people's ability to infer association or independence between two variables given either a sequence of cases, or a summary table of the data (De Houwer and Beckers (2002) review empirical data on contingency learning; McKenzie and Mikkelsen (2007) review formal models). Three different models of asymptotic causal inference give a feel for their diversity. The ΔP model (Cheng and Novick 1990; 1992) holds that causal strength judgements are given by the difference between the probability of the effect when (1) the potential cause is

present, and (2) when it is absent (i.e. $\mathit{\Delta}P=P(E|C)-P(E|\neg C)$). The causal power approach (Cheng 1997; Novick and Cheng 2004) supposes that people represent the world in terms of unobserved causal powers (similar to capacities in the sense of Cartwright 1989) and use the observed statistics to try to make inferences about the strengths of those powers. The pCI model (White 2003*a*; 2003*c*) argues that people attend to the proportion of confirming instances: the fraction of observed cases that support the existence of a causal relation (relative to the total number of observed cases). There are differences in metaphysical commitments and mathematics, but the models of asymptotic causal inference all share the common goal of predicting people's rating patterns once learning is completed and beliefs have stabilized.

Associative models of causal inference and models of asymptotic causal inference are often thought to be direct competitors with one another. A series of formal results (e.g. Cheng 1997; Danks 2003; Tenenbaum and Griffiths 2001) have emerged in the past ten years, however, showing systematic connections between associative models of causal inference and models of asymptotic causal inference. Specifically, many different associative models of case-by-case learning each (provably) converge in the limit to a different asymptotic model (e.g. the associative model of Danks et al. (2003) converges to causal power). Danks (2007*b*) extends and unifies these disparate results, and shows that these different types of models do not compete, but rather are simply models at different temporal scales.

These mathematical connections reveal that both associative and asymptotic models focus on inference of causal strengths, rather than causal structure. Of course, strength ratings implicitly encode structure (i.e. no causal connection if and only if strength of zero), but the two types of inference are at least logically separable. The causal Bayes net framework (Pearl 2000; Spirtes, Glymour, and Scheines 1993) explicitly represents this distinction between causal structure (i.e. the graph) and causal strength (i.e. the parameters). Moreover, all the previously proposed causal inference theories—both associative and asymptotic—provably correspond to strength inference rather than structure inference (Danks 2007*b*; Griffiths and Tenenbaum 2005). The causal Bayes net framework also provides a clear account of the difference between learning from observations and learning from interventions, which had previously been largely neglected in the psychological literature. That account led to numerous studies that confirmed that the observation vs. intervention difference affects people's causal inference (e.g. Gopnik et al. 2004; Lagnado and Sloman 2004; Sobel and Kushnir 2006; Steyvers et al. 2003).

This representational power, as well as the successful use of causal Bayes nets in other domains, has prompted two types of proposal for human causal inference based on causal Bayes nets. The first type holds that human causal inference involves learning causal structure and strengths from the set of all possible structures consistent with background knowledge (Gopnik and Glymour 2002; Gopnik et al. 2004; Griffiths and Tenenbaum 2005; Steyvers et al. 2003). These proposals differ principally about the nature and level of the learning algorithm. The second type argues that people start with some initial structure, and only change their mind if the data directly contradict the initial model (Hagmayer et al. 2007; Lagnado and Sloman 2004; Lagnado et al. 2007; Waldmann 1996). The initial structure is selected on the basis of various heuristics, such as 'if I change *X*'s value, then anything that changes afterwards must be an effect of *X*'. Learning on this account does not involve selecting the

best (by some measure) causal Bayes net from the set of all plausible possibilities; instead, the learner selects a hypothesis by various heuristics, and then retains it until it is falsified.

There are many different psychological models for causal inference, and a correspondingly large number of experimental studies. There is very little agreement about which causal inference model is right, or even about which approach is the most likely to be fruitful. One significant problem is that none of the extant models can capture all of the extant data (Perales and Shanks 2007). Some models are, of course, better than others, but none predict all the ways that ratings vary as statistical information changes. Causal Bayes net theories have a representational advantage in capturing the psychologically significant distinction between observation and intervention, but even they cannot explain all data relating to that distinction. One potential explanation is that people use a mixture of strategies (Buehner, Cheng, and Clifford 2003; Lober and Shanks 2000), and different experiments may well elicit different types of judgement, as ratings seem to be sensitive to the probe question used (Collins and Shanks 2006; White 2003*b*). One might hope that neuroscience could help, but currently available fMRI data do not provide much insight. Causal inference does seem to involve some sort of error prediction/correction occurring principally in the prefrontal cortex (Corlett et al. 2004; Fletcher et al. 2001; Turner et al. 2004), but the data tell us nothing about how that calculation figures in causal inference. Semantic retrieval of causal information and semantic retrieval of associative information lead to different patterns of neural activation (Satpute et al. 2005), but again the data do not illuminate the nature of that difference.

The preceding discussion has largely focused on a limited subset of causal inference: principally, causal inference from statistical information, rather than other information. In practice, causal inference is sensitive to many other factors, such as knowledge of temporal features of the possible causal relations (Buehner and May 2002; 2003; Lagnado and Sloman 2004). More generally, causal inference is clearly influenced by prior beliefs. Strong co-variations in observed data are more meaningful (i.e. lead to larger ratings) if people know a plausible mechanism underlying the co-variation, rather than an implausible one; relatively little effect of mechanism plausibility in prior belief occurs for weak co-variations, perhaps because ratings are already quite low (Fugelsang and Thompson 2003). This effect also seems to have a neural basis. Consistency between prior belief and observed data (i.e. plausible mechanism and strong co-variation, or implausible mechanism and weak co-variation) activates learning and memory regions of the brain, while belief-data inconsistency activates error-correction and conflict resolution areas (Fugelsang and Dunbar 2005). Most generally, causal inference is significantly influenced by the categories and concepts that we have (Waldmann and Hagmayer 2006). People typically do causal inference with the categories that they have prior to learning, even when those categories are suboptimal for causal inference.

4. Intersections between Causal Perception and Causal Inference

An obvious issue centres on the relationship, if any, between causal perception and causal inference. Are the processes identical? Is one necessary for the other? Is one a subset of the other? Are they entirely distinct? One way to (try to) address this issue is through developmental progressions; for example, if one type of cognition appears before the other, then the later process presumably cannot be necessary for the earlier one. As noted earlier,

causal perception has been found in 6-montholds. It is at least prima facie possible that infants of a similar age make causal inferences, particularly since at least 8-month-olds are sensitive to some statistical patterns in their environment (Saffran, Aslin, and Newport 1996; Saffran et al. 1999). We do not know, however, the earliest age of causal inference, largely because there are obvious methodological challenges. There will typically be many alternative explanations for data from looking-time studies that suggest causal inference, such as the infants simply noticing predictively useful associations. We thus do not currently know whether one process emerges before the other.

A different approach is to explore whether the processes can be separated in adult cognition. Surprisingly, there has been relatively little research directly on this topic. On the neurological front, Roser et al. (2005) examined causal perception and inference in two corpus callosotomy[4] patients, and found that causal perception and causal inference seem to occur in different brain hemispheres (perception in right hemisphere, inference in left hemisphere). Independent fMRI studies on each type of causal learning also suggest a neuroanatomical difference. Causal perception seems to be concentrated in the temporal lobes (Blakemore et al. 2001; Fugelsang et al. 2005), while at least one significant part of causal inference—namely, error prediction and correction—seems to be largely localized in the prefrontal cortex (Corlett et al. 2004; Fletcher et al. 2001; Turner et al. 2004). However, despite these apparent neuroanatomical differences between causal inference and perception, there are no known cases of selective loss of only one of the types of cognition. The neurological evidence is thus relatively ambiguous: causal perception and causal inference seem to occur at least partially in different brain regions, but it is unknown whether they are fully dissociable.

Yet another approach is to try to find situations that prompt only causal perception or only causal inference. Schlottmann and Shanks (1992) presented experimental participants with many different sets of launching-type sequences. In one set of sequences, spatio-temporal contact reliably led to movement of the 'launched' block only after a noticeable delay. Participants came to recognize, presumably via causal inference, that the 'launching' block was a cause, but they reported that 'it just did not look as if it should be' (ibid. 338) and so gave relatively low causal perception ratings. In a second set of sequences, spatio-temporal contact was uncorrelated with subsequent launching and instead, colour change in the launched block was the reliable predictor. The most interesting case for these sequences is when the second block moves after spatio-temporal contact, even though such contact is (overall) uncorrelated with movement. In this case, participants give high causal perception ratings for the 'launching' block as a cause, even though they recognize that it is entirely unnecessary; they report that the collision 'just looked as if it should be' a cause (ibid. 338). This distinction between causal perception and causal inference is found in both participant ratings and phenomenological experiences (ibid. 339), which supports the idea that these types of cognition are actually separable.

Interactions between causal perception and causal inference can become quite complicated when a causal mechanism requires some time to operate. For example, suppose a button press causes a light to illuminate only (and always) after a three-second interval. Now suppose that one presses that button, and then presses it again three seconds later (when the light comes on). Causal perception says the second button press is the cause because of temporal

contiguity; causal inference (or reasoning) says the first button press is the cause. Adults are largely able to use mechanism information to override the causal perception in these cases, but 7-yearold children are not (Schlottmann 1999). More generally, adult causal inference is influenced by knowledge of the timing of underlying mechanisms, as cause–effect relationships can be inferred even when there is a significant temporal gap between the two (Buehner and May 2002; 2003). Adult causal perception is not influenced by that knowledge, however, as spatio-temporally contiguous events are (almost) always perceived causally while separated ones are not. Explicit timing knowledge in causal inference can also shape which events are thought to be possible causes in the first place (Hagmayer and Waldmann 2002), but has no such impact on causal perception.

In summary, there is a growing body of direct and indirect evidence that causal perception and causal inference are different cognitive processes. The current psychological evidence does not, however, provide much information about the relationship between these processes. In particular, it is simply unknown whether one is necessary for the other—either developmentally or cognitively—or they are (relatively) autonomous cognitive processes. One barrier to fruitful psychological research has arguably been the lack of understanding of the relevant theoretical 'possibility space'. The space of possible relationships between causal perception and causal inference is largely unknown, and philosophical thought could potentially provide significant guidance in the development and testing of psychological theories.

5. CAUSAL REASONING

Causal reasoning is also a significant part of causal cognition, and perhaps even constitutes the majority of adult causal cognition. One great value of causal knowledge is the myriad ways that we can use it to understand, predict, and control the world around us (Sloman 2005). Psychological research on human causal reasoning has historically been conducted in particular domains of application of the causal knowledge (e.g. decision-making, categorization). In recent years, the causal Bayes net formalism has provided a small measure of unification to the reasoning research, but the work still largely consists of various disjoint research endeavours.

One commonplace type of causal reasoning is the use of causal knowledge to make decisions. Suppose I know (or believe) that *X* causes *Y*. If I desire *Y*, then I might naturally decide to try to bring about *X*. In contrast, if I desire *X*, then there is no particular value to bringing about *Y* directly. People are sensitive to this distinction, and they exhibit appropriate behaviour whether they are taught the causal structure explicitly or learn it from observed data (Hagmayer and Sloman 2005; Nichols and Danks 2007). People also appear systematically to treat their own decisions as occurring outside the causal system; they act (except in very unusual situations) as though their decisions are uncaused by variables in the causal structure. Moreover, many of the experiments used to explore causal inference employ behavioural measures of learning that depend on people's ability to do causal reasoning. For example, children are shown that some combinations of blocks ('blickets') activate a machine, and then their causal knowledge is assessed by asking them to make the machine go or stop (Gopnik et

al. 2004). This behavioural measure—which block the child places on the detector, or removes from it—is discriminative of causal learning only if the children are able to use the products of learning to make decisions. This range of findings about decision-making based on causal reasoning has led to a formal model of decision-making (given causal beliefs) that is based on causal Bayes nets (Sloman and Hagmayer 2006), and experimental tests are ongoing.

Causal reasoning also occurs in the context of conceptual reasoning. One example of the relevance of causal beliefs to categories comes from the 'causal status effect'. If category A is partially characterized by the (possibly indeterministic) causal relation $X \rightarrow Y$, then individuals with X but not Y are systematically judged as more likely to be in A than individuals with Y but not X (Ahn et al. 2000; Rehder and Kim 2006). That is, if all As have X causing Y, then X is more important than Y in deciding whether some new individual is an A. Rehder and colleagues have argued that the connection between causal reasoning and concepts might be substantially deeper. In the past forty years, psychological theories of concepts have usually understood concepts in terms of observed features, whether prototypical instances (Posner and Keele 1968), sets of exemplars (Nosofsky 1984), sets of typical features (Tversky 1977), or something else. A different idea is that at least some categories might be defined by shared causal structure: two objects fall under the same concept if and only if they have the same underlying causal structure (Rehder 2003a; 2003b; Rehder and Kim 2006). Specifically, the causal model theory holds that some concepts are defined by a common causal structure, almost always expressed as a causal Bayes net, and that conceptual reasoning is essentially causal reasoning. For example, causal model theory holds that similarity judgments—how similar some new object O is to category A—are given by $P(O|A)$: the probability that an object randomly chosen from A would be like O (Rehder 2003b), and then those similarities are used to produce categorization judgements (i.e. $P(A|O)$).[5] Feature inference—given that object O of type A has features F_1, F_2, etc., how likely it is that O has feature G—is similarly understood as causal reasoning: probabilistic inference in a particular causal Bayes net given observations of some of the variables (Rehder and Hastie 2004). Causal model theory has led to numerous experimental results that demonstrate clearly the importance of causal beliefs and reasoning in people's concepts, though substantial open questions remain (e.g. its scope of applicability, and its ability to represent conceptual hierarchies).

Causal reasoning is also closely connected to counterfactual reasoning, as our causal knowledge often plays a role in assessing counterfactuals, and counterfactual 'but for' reasoning is frequently part of causal reasoning. Psychological research on counterfactual reasoning has focused on both evaluation of the truth of particular counterfactuals, and on the spontaneous generation of counterfactuals (Mandel, Hilton, and Catellani 2005). Suppose that factors C_1, ..., C_n, and outcome E all occur. People are more likely to judge the counterfactual 'If not-C_i, not-C_j ..., then not-E' as true to the extent that the factors in the antecedent (1) are anomalous; (2) are controllable; (3) violate a social or moral norm; (4) are close in time or space to the outcome; and/or (5) have a known mechanism connecting them to E. The same dimensions seem to be relevant for which factors are mentioned in the antecedent of spontaneously generated counterfactuals (Byrne 2005; Roese 1997 provide reviews). One clear conclusion is that, although causal and counterfactual reasoning are closely connected, they are not identical with one another (Mandel 2003). As an example, consider an individual who

takes an unusual route home, but is hit by a drunk driver during the drive. When asked to think about relevant counterfactuals for this individual, most people respond: 'if she hadn't taken the unusual route, then she wouldn't have been involved in the accident'. At the same time, most people judge the drunk driver to be the principal cause of the accident. More generally, the factors that are weighted most heavily in counterfactual reasoning are not necessarily the ones that are judged either to have the greatest causal influence or to be the most causally relevant. Counterfactual and causal reasoning make use of each other, but are not the same cognitive process.

These previous lines of research all provided indirect ways to study causal reasoning. A direct approach is to study people's causal reasoning when they have to determine the causes of some token event. For example, people presumably decide which event in some sequence caused a car accident by causal reasoning about their prior beliefs. Psychological research on this problem of causal attribution has primarily focused on whether people use knowledge of mechanisms or of correlations to make these decisions (Ahn and Bailenson 1996; Ahn et al. 1995). Suppose, for example, that I know that taking some medication is correlated with car accidents, and I know a mechanism by which wet roads lead to car accidents. When asked about the cause of a car accident involving both the medication and a wet road, I am more likely to attribute the accident to the wetness of the road. In general, people prefer to use mechanism information, and when both types of information are used, they weight mechanism information more heavily (Ahn and Bailenson 1996; Ahn et al. 1995). This result is not particularly surprising, as knowledge of a mechanism usually implies knowledge of when it is (and is not) likely to be active; knowledge of correlations often does not have the same type of scope knowledge. This dependence on mechanism knowledge in causal attribution is particularly striking, however, given that most people have an 'illusion of explanatory depth' (Rozenblit and Keil 2002): an overestimation of their own mechanism knowledge and difficulty accepting that their knowledge is limited in this regard. In general, people believe that they can explain the mechanism M underlying $X \rightarrow Y$, and then explain the mechanisms underlying M, and so on; actually, they can rarely describe anything more than M. Importantly, however, the notion of 'mechanism' used in this research is much broader and weaker than the notion recently advanced in the philosophy of science (e.g. Craver 2007; Machamer, Darden, and Craver 2000). The mechanism information in Ahn et al.'s studies principally consists of intermediate events or variables, rather than any knowledge of how the pieces fit together (Danks 2005).

6. CAUSALITY IN NON-HUMAN ANIMALS

A final source of information about the psychology of causation comes from comparative studies with other animal species. Historically, animal research has focused on classical and instrumental conditioning (described in sect. 3). Non-human animals can certainly predict future events, but they were not thought to have any substantive notion of causation in the world; it was assumed that the predictions were based entirely on learned associations, or relatively domain-specific triggers (e.g. Lavin, Freise, and Coombes 1980). This consensus view then shifted, particularly with respect to non-human primates, as numerous field studies emerged that reported extensive and seemingly sophisticated tool use by wild animals. For

example, chimpanzees were found to 'fish' for termites by inserting long, flexible implements (grass, reeds, sticks, etc.) into termite mounds, waiting for the termites to latch onto the implement, and then removing it for a tasty termite snack (Goodall 1986). Chimpanzees were even observed to modify their tools in ways that improved their performance. These observations led many authors to argue that some non-human animals have a relatively rich notion of causality, and perhaps even the same concept as humans (McGrew 1992; Premack 1976). These claims were principally about non-human primates, though there have periodically been similar claims about other non-human animals.

In recent years, however, the view that non-human animals—primates in particular—have a rich, almost-human notion of causation has come under increased attack. Tomasello and Call (1997) surveyed a wide range of primate behaviours and argued that non-human primates have only very rich associations, rather than any notion of causality that is abstract, or domain-general, or involves unobserved forces. That is, perhaps the remarkably complex behaviour exhibited by non-human primates arises from remarkably complex, but entirely perception-based, associations between object shapes, action sequences, and outcomes. Povinelli and colleagues have carried out a number of experiments (summarized in Povinelli 2000) that explicitly try to determine if a chimpanzee's learning involves abstract causal knowledge. Many of their experiments find that chimpanzees are seemingly insensitive to the underlying causal structure of a situation, and respond only to superficial perceptual cues. For example, chimpanzees seem to think that a rope that is touching an object can always be used to pull that object towards them; they do not seem to be sensitive to the fact that a physical *connection* is required, not merely physical contact (Povinelli 2000: ch. 9). If given enough trials, then chimpanzees would of course 'learn' the difference between a rope on top of a banana and one tied around the banana, but only through repeated associations between various perceptions (rope on top vs. rope tied around) and success or failure in obtaining the banana. The chimpanzee's eventual success would not be based on reasoning or learning with an abstract, domain-general notion of causation that requires physical connection to manifest. More generally, one can argue that essentially all findings of 'causal learning' or 'causal reasoning' in non-human animals can be explained in similar ways (Penn and Povinelli 2007).

These findings have been used to argue for a positive hypothesis about the uniqueness of human behaviour (Penn and Povinelli 2007; Penn, Holyoak, and Povinelli 2008; Povinelli 2000). The 'reinterpretation hypothesis' holds that only humans are able to reinterpret the surface features of the world in terms of complex, necessarily unobserved predicates and relations (e.g. causality, support, force, etc.). In particular, it holds that only humans can take a sequence of perceptual inputs, and explain or describe that sequence in terms of causality. The reinterpretation hypothesis essentially says that Hume's problem never arises for non-human animals, since they never (re-)conceptualize the world in terms of unobserved causal influences. Of course, non-human animals have concepts of varying complexity, but the reinterpretation hypothesis argues that those concepts are always restricted to the perceptual.

Much of the work on causal learning and reasoning in humans points towards us having a complex, multifaceted concept, or perhaps even multiple concepts of causation (separate for causal perception and causal inference). Even if we have only a single concept of causation, it surely involves many different dimensions, properties, and relations. We must therefore take

care that our understanding of the notion of causation in non-human animals is not based on some simple dichotomy of associationism vs. full-blooded causal learning/reasoning. We should take seriously the possibility that, even if non-human animals do not have the same notion of causality as humans, they need not be 'mere' associationists. Two recent experiments with rats illustrate the large middle ground between the endpoints of the standard, simplistic dichotomy.

One standard tenet of associationism is that observations involving a cue (e.g. a tone) affect only the associative strengths of that cue, and perhaps also the strengths of other cues that have reliably co-occurred with that cue in the past. If a tone has never occurred with a light, for example, then further observations of the tone should not (on the standard account) affect associations involving the light. This assumption turns out to be false, as rats use observations of one cue to revise associative strengths of cues that have never co-occurred with the original cue (Denniston et al. 2003).[6] The rats' behaviour suggests that they are using some sort of higher-order 'reasoning' about the relationships between the various cues, though the exact nature of that reasoning is currently unknown. At the very least, the 'associationist' processes of rats are substantially more sophisticated than the standard account suggests.

An even more striking finding comes from Blaisdell et al. (2006), who have recently argued that rats seem to do causal reasoning using (something like) causal Bayes nets. Consider two different causal structures: (a) $X \leftarrow Y \rightarrow Z$; and (b) $X \rightarrow Y \rightarrow Z$. Blaisdell et al. (2006) split their rats into two groups, and used classical conditioning to 'teach' each group both edges in one of the causal structures. For example, group (a) received repeated trials of Y followed by X, and separate trials of Y followed by Z;in both groups, X = tone; Y = light; and Z = food. Now consider an intervention to bring about X. In the common-cause structure (a), the intervention will break the $X Y$ causal connection, and so neither Y nor Z will change; in the chain structure (b), the intervention will not change the causal structure, and so both Y and Z will (probabilistically) change. Both groups of rats were provided with such an 'intervention' in the form of a lever that produced the tone X. Rats that 'learned' the common-cause structure were significantly less likely to check for the food after pressing the lever, compared to rats that 'learned' the chain structure. That is, the rats behaved as if they knew whether their actions to produce X were likely to bring about a change in Z, and they acted consistently with the predictions of a causal Bayes net model. This finding does not, of course, prove that rats have causal Bayes nets 'in their heads'. There are more minimal interpretations of this result, and prediction given interventions is only one aspect of causal Bayes nets (Penn and Povinelli 2007). This finding does show, however, that rats are more than just simple associationists: they seem to be able to integrate distinct pieces of observational evidence into a single, relatively coherent structure, and then use that observational evidence to make predictions about the outcomes of interventions (see also Leising et al. 2008).

Causal learning and reasoning in humans—the full array, scope, and types—seem to be unique among the animal kingdom, and the nature and source of this uniqueness is the subject of ongoing debate (e.g. Penn, Holyoak, and Povinelli 2008 and accompanying commentaries). Non-human animals are capable of remarkably sophisticated behaviour that takes advantage of causal relations in the world to help them reach their goals, but seemingly always in ways that differ crucially, though not always obviously, from human behaviour. Research on non-human animals is nonetheless potentially able to provide some insight into human causal learning and

reasoning precisely by revealing the multifaceted nature of our causal concept(s) and cognition. The ways in which non-human animals are more than simple associationists potentially indicate some of the components of causal cognition in humans. One significant open question is how to use philosophical research on the many dimensions and uses of causation to inform research on which components of human causal cognition are found in non-human animals.

7. CONCLUSION

Psychological research on causation has expanded rapidly in the last twenty years, and it seems to be one of the 'hot areas' in cognitive science right now. In these years, there have been significant theoretical and empirical advances on causal perception, inference, and reasoning, though many open questions remain in all three areas. One striking feature of this research is that it has almost all focused on causal cognition in isolation. For example, an experiment will ask participants to learn some causal structure from observed cases, but then ask only for explicit, verbal causal strength judgements. People are almost never asked to make meaningful decisions using the causal information that they learn. In contrast with these laboratory experiments, causal cognition 'in the wild' cannot easily be isolated from other cognitive processes. One of the most significant challenges facing psychologists in coming years is thus to understand better the relationship between causal cognition and other cognitive processes, such as decision-making, linguistic inferences/pragmatics, and social behaviour.

FURTHER READING

Sloman (2005): Describes many of the ways that causal knowledge and reasoning are relevant to other parts of cognition. Provides an overview of causal Bayes nets and defends them as a psychological account of causal knowledge.

Michotte (1963): Classic text on causal perception. Includes a wide range of experiments on the launching effect.
Scholl and Tremoulet (2000): More recent review article on the launching effect. Argues that causal perception is modular (or close to it).
Gopnik et al. (2004): Major paper defending the causal Bayes net view of causal representation and learning. Describes much of the developmental evidence about causal learning.
Danks (2007*b*): Provides an overview and unification of all the major theories of causal inference.
Gopnik and Schulz (2007): Edited book covering both causal inference and causal reasoning. Most papers either use, or respond to theories that use, causal Bayes nets.
Sloman and Hagmayer (2006): Presents a theory of decision-making based on causal knowledge, represented as causal Bayes nets.
Mandel (2003): Surveys much of the empirical data about causal and counter-factual reasoning, and argues that they are importantly distinct.

Povinelli (2000): Describes many of the experiments that expose the limits of causal knowledge in non-human primates. Carefully explores methodological challenges facing research on animal cognition.

Penn and Povinelli (2007): Critical review of research on causal cognition in non-human animals. Questions whether non-human animals have a rich notion of causation.

REFERENCES

AHN, W.-K., and BAILENSON, J. (1996). 'Causal Attribution as a Search for Underlying Mechanisms: An Explanation of the Conjunction Fallacy and the Discounting Principle', *Cognitive Psychology* 31: 82–123.

—— KALISH, C. W., MEDIN, D. L., and GELMAN, S. A. (1995). 'The Role of Covariation Versus Mechanism Information in Causal Attribution', *Cognition* 54: 299–352.

—— KIM, N. S., LASSALINE, M. E., and DENNIS, M. J. (2000). 'Causal Status as a Determinant of Feature Centrality', *Cognitive Psychology* 41: 361–416.

BEASLEY, N. A. (1968). 'The Extent of Individual Differences in the Perception of Causality', *Canadian Journal of Psychology* 22: 399–407.

BLAISDELL, A. P., SAWA, K., LEISING, K. J., and WALDMANN, M. R. (2006). 'Causal Reasoning in Rats', *Science* 311: 1020–2.

BLAKEMORE, S.-J., FONLUPT, P., PACHOT-CLOUARD, M., DARMON, C., BOYER, P., MELTZOFF, A. N., SEGEBARTH, C., and DECETY, J. (2001). 'How the Brain Perceives Causality: An Event-Related fMRI Study', *NeuroReport* 12: 3741–6.

BUEHNER, M. J., and MAY, J. (2002). 'Knowledge Mediates the Timeframe of Covariation Assessment in Human Causal Induction', *Thinking & Reasoning* 8: 269–95.

— — (2003). 'Rethinking Temporal Contiguity and the Judgement of Causality: Effects of Prior Knowledge, Experience, and Reinforcement Procedure', *Quarterly Journal of Experimental Psychology* 56A: 865–90.

—— CHENG, P. W., and CLIFFORD, D. (2003). 'From Covariation to Causation: A Test of the Assumption of Causal Power', *Journal of Experimental Psychology: Learning, Memory, & Cognition* 29: 1119–40.

BYRNE, R. M. J. (2005). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, Mass.: MIT.

CARTWRIGHT, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

CATENA, A., MALDONADO, A., and CANDIDO, A. (1998). 'The Effect of the Frequency of Judgment and the Type of Trials on Covariation Learning', *Journal of Experimental Psychology: Human Perception and Performance* 24: 481–95.

CHENG, P. W. (1997). 'From Covariation to Causation: A Causal Power Theory', *Psychological Review* 104: 367–405.

—— and NOVICK, L. R. (1990). 'A Probabilistic Contrast Model of Causal Induction', *Journal of Personality and Social Psychology* 58: 545–67.

— — (1992). 'Covariation in Natural Causal Induction', *Psychological Review* 99: 365–82.

CHOI, H., and SCHOLL, B. J. (2004). 'Effects of Grouping and Attention on the Perception

of Causality', *Perception & Psychophysics* 66: 926–42.

— — (2006). 'Perceiving Causality after the Fact: Postdiction in the Temporal Dynamics of Causal Perception', *Perception* 35: 385–99.

COHEN, L. B., and AMSEL, G. N. (1998). 'Precursors to Infants' Perception of the Causality of a Simple Event'. *Infant Behavior and Development* 21: 713–31.

COLLINS, D. J., and SHANKS, D. R. (2006). 'Conformity to the Power PC Theory of Causal Induction Depends on the Type of Probe Question', *Quarterly Journal of Experimental Psychology* 59: 225–32.

CORLETT, P. R., AITKEN, M. R. F., DICKINSON, A., SHANKS, D. R., HONEY, G. D., HONEY, R. A. E., ROBBINS, T. W., BULLMORE, E. T., and FLETCHER, P. C. (2004). 'Prediction Error During Retrospective Revaluation of Causal Associations in Humans: fMRI Evidence in Favor of an Associative Model of Learning', *Neuron* 44: 877–88.

CRAVER, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Oxford University Press.

CSIBRA, G., GERGELY, G., BÍRÓ, S., KOÓS, O., and BROCKBANK, M. (1999). 'Goal Attribution without Agency Cues: The Perception of "Pure Reason" in Infancy', *Cognition* 72: 237–67.

DANKS, D. (2003). 'Equilibria of the Rescorla-Wagner Model', *Journal of Mathematical Psychology* 47: 109–21.

—— (2005). 'The Supposed Competition between Theories of Human Causal Inference', *Philosophical Psychology* 18: 259–72.

—— (2007*a*). 'Theory Unification and Graphical Models in Human Categorization', in A. Gopnik and L. E. Schulz (eds.), *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press, 173–89.

—— (2007*b*). 'Causal Learning from Observations and Manipulations', in M. C. Lovett and P. Shah (eds.), *Thinking with Data*. Mahwah, NJ: Lawrence Erlbaum, 359–88.

—— GRIFFITHS, T. L., and TENENBAUM, J. B. (2003). 'Dynamical Causal Learning', in S. Becker, S. Thrun, and K. Obermayer (eds.), *Advances in Neural Information Processing Systems* 15. Cambridge, Mass.: MIT, 67–74.

DE HOUWER, J., and BECKERS, T. (2002). 'A Review of Recent Developments in Research and Theories on Human Contingency Learning', *Quarterly Journal of Experimental Psychology* 55B: 289–310.

DENNISTON, J. C., SAVASTANO, H. I., BLAISDELL, A. P., and MILLER, R. R. (2003). 'Cue Competition as a Retrieval Deficit", *Learning and Motivation* 34: 1–31.

FLETCHER, P. C., ANDERSON, J. M., SHANKS, D. R., HONEY, R. A. E., CARPENTER, T. A., DONOVAN, T., PAPADAKIS, N., and BULLMORE, E. T. (2001). 'Responses of Human Frontal Cortex to Surprising Events Are Predicted by Formal Associative Learning Theory', *Nature Neuroscience* 4: 1043–8.

FODOR, J. A. (1983). *The Modularity of Mind*. Cambridge, Mass.: MIT.

FONLUPT, P. (2003). 'Perception and Judgement of Physical Causality Involve Different Brain Structures', *Cognitive Brain Research* 17: 248–24.

FUGELSANG, J. A., and DUNBAR, K. N. (2005). 'Brain-Based Mechanisms Underlying Complex Causal Thinking', *Neuropsychologia* 42: 1204–13.

—— and THOMPSON, V. A. (2003). 'A Dual-Process Model of Belief and Evidence

Interactions in Causal Reasoning', *Memory & Cognition* 31: 800–15.

—— ROSER, M. E., CORBALLIS, P. M., GAZZANIGA, M. S., and DUNBAR, K. N. (2005). 'Brain Mechanisms Underlying Perceptual Causality', *Cognitive Brain Research* 24.

GERGELY, G., NÁDASDY, Z., CSIBRA, G., and BÍRÓ, S. (1995). 'Taking the Intentional Stance at 12 Months of Age'. *Cognition*, 56: 165–93.

GOLDVARG, E., and JOHNSON-LAIRD, P. N. (2001). 'Naive Causality: A Mental Model Theory of Causal Meaning and Reasoning', *Cognitive Science* 25: 565–610.

GOODALL, J. (1986). *The Chimpanzees of Gombe: Patterns of Behavior*. Cambridge, Mass.: Harvard University Press.

GOPNIK, A., and GLYMOUR, C. (2002). 'Causal Maps and Bayes Nets: A Cognition and Computational Account of Theory-Formation', in P. Carruthers, S. Stich, and M. Siegal (eds.), *The Cognitive Basis of Science*. Cambridge: Cambridge University Press, 117–32.

—— and SCHULZ, L. E. (eds.) (2007). *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press.

—— GLYMOUR, C., SOBEL, D. M., SCHULZ, L. E., KUSHNIR, T., and DANKS, D. (2004). 'A Theory of Causal Learning in Children: Causal Maps and Bayes Nets', *Psychological Review* 111: 3–32.

GRIFFITHS, T. L., and TENENBAUM, J. B. (2005). 'Structure and Strength in Causal Induction', *Cognitive Psychology* 51: 334–84.

HAGMAYER, Y., and SLOMAN, S. A. (2005). 'A Causal Model Theory of Choice', in B. G. Bara, L. Barsalou, and M. Bucciarelli (eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum, 881–6.

—— and WALDMANN, M. R. (2002). 'How Temporal Assumptions Influence Causal Judgments', *Memory & Cognition* 30: 1128–37.

—— SLOMAN, S. A., LAGNADO, D. A., and WALDMANN, M. R. (2007). 'Causal Reasoning through Intervention', in A. Gopnik and L. E. Schulz (eds.), *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press, 86–100.

HEIDER, F., and SIMMEL, M.-A. (1944). 'An Experimental Study of Apparent Behavior', *American Journal of Psychology* 57: 243–9.

LAGNADO, D. A., and SLOMAN, S. A. (2004). 'The Advantage of Timely Intervention', *Journal of Experimental Psychology: Learning, Memory, & Cognition* 30: 856–76.

—— WALDMANN, M. R., HAGMAYER, Y., and SLOMAN, S. A. (2007). 'Beyond Covariation: Cues to Causal Structure', in A. Gopnik and L. E. Schulz (eds.), *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press, 154–72.

LAVIN, M. J., FREISE, B., and COOMBES, S. (1980). 'Transferred Flavor Aversions in Adult Rats', *Behavioral and Neural Biology* 28: 15–33.

LEISING, K. J., WONG, J., WALDMANN, M. R., and BLAISDELL, A. P. (2008). 'The Special Status of Actions in Causal Reasoning in Rats', *Journal of Experimental Psychology: General* 137/3: 514–27

LESLIE, A. M. (1982). 'The Perception of Causality in Infants', *Perception* 11: 173–86.

—— (1984). 'Spatiotemporal Continuity and the Perception of Causality in Infants', *Perception* 13: 287–305.

—— (1994). 'ToMM, ToBy, and Agency: Core Architecture and Domain Specificity', in

L. Hirschfield and S. A. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press, 119–48.

—— and KEEBLE, S. (1987). 'Do Six-Month-Old Infants Perceive Causality?' *Cognition* 25: 265–88.

LOBER, K., and SHANKS, D. R. (2000). 'Is Causal Induction Based on Causal Power? Critique of Cheng (1997)', *Psychological Review* 107: 195–212.

MCGREW, W. C. (1992). *Chimpanzee Material Culture: Implications for Human Evolution*. Cambridge: Cambridge University Press.

MACHAMER, P., DARDEN, L., and CRAVER, C. F. (2000). 'Thinking About Mechanisms', *Philosophy of Science* 67: 1–25.

MCKENZIE, C. R. M., and MIKKELSEN, L. A. (2007). 'A Bayesian View of Covariation Assessment', *Cognitive Psychology* 54: 33–61.

MANDEL, D. R. (2003). 'Judgment Dissociation Theory: An Analysis of Differences in Causal, Counterfactual, and Covariational Reasoning', *Journal of Experimental Psychology: General* 132: 419–34.

—— HILTON, D. J., and CATELLANI, P. (eds.) (2005). *The Psychology of Counterfactual Thinking*. New York: Routledge.

MICHOTTE, A. (1963). *The Perception of Causality*. London: Methuen.

MORRIS, M. W., and PENG, K. (1994). 'Culture and Cause: American and Chinese Attributions for Social and Physical Events'. *Journal of Personality and Social Psychology* 67: 949–71.

NEWMAN, G. E., CHOI, H., WYNN, K., and SCHOLL, B. J. (2008). 'The Origins of Causal Perception: Evidence from Postdictive Processing in Infancy', *Cognitive Psychology* 57/3: 262–91.

NICHOLS, W., and DANKS, D. (2007). 'Decision Making Using Learned Causal Structures', in D. S. McNamara and J. G. Trafton (eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*. Austin, Tex.: Cognitive Science Society, 1343–8.

NOSOFSKY, R. M. (1984). 'Choice, Similarity, and the Context Theory of Classification', *Journal of Experimental Psychology: Learning, Memory, & Cognition* 10: 104–14.

NOVICK, L. R., and CHENG, P. W. (2004). 'Assessing Interactive Causal Influence', *Psychological Review* 111: 455–85.

OAKES, L. M. (1994). 'Development of Infants' Use of Continuity Cues in Their Perception of Causality', *Developmental Psychology* 30: 869–79.

—— and COHEN, L. B. (1990). 'Infant Perception of a Causal Event', *Cognitive Development* 5: 193–207.

PEARCE, J. M. (1987). 'A Model for Stimulus Generalization in Pavlovian Conditioning', *Psychological Review* 94: 61–73.

PEARL, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

PENN, D. C., and POVINELLI, D. J. (2007). 'Causal Cognition in Humans and Nonhuman Animals: A Comparative, Critical Review', *Annual Review of Psychology* 58: 97–118.

—— HOLYOAK, K. J., and POVINELLI, D. J. (2008). 'Darwin's Mistake: Explaining the Discontinuity between Human and Nonhuman Minds', *Behavioral and Brain Sciences* 31: 109–30.

PERALES, J. C., and SHANKS, D. R. (2007). 'Models of Covariation-Based Causal Judgment: A Review and Synthesis', *Psychonomic Bulletin & Review* 14: 577–96.

POSNER, M. I., and KEELE, S. W. (1968). 'On the Genesis of Abstract Ideas', *Journal of Experimental Psychology* 77: 353–63.

POVINELLI, D. J. (2000). *Folk Physics for Apes: The Chimpanzee's Theory of How the World Works*. Oxford: Oxford University Press.

PREMACK, D. (1976). *Intelligence in Ape and Man*. Hillsdale, NJ: Erlbaum.

REHDER, B. (2003*a*). 'A Causal-Model Theory of Conceptual Representation and Categorization', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 1141–59.

—— (2003*b*). 'Categorization as Causal Reasoning', *Cognitive Science* 27: 709–48.

—— and HASTIE, R. (2004). 'Category Coherence and Category-Based Property Induction', *Cognition* 91: 113–53.

—— and KIM, S. (2006). 'How Causal Knowledge Affects Classification: A Generative Theory of Categorization', *Journal of Experimental Psychology: Learning, Memory, & Cognition* 32: 659–83.

RESCORLA, R. A., and WAGNER, A. R. (1972). 'A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement', in A. H. Black and W. F. Prokasy (eds.), *Classical Conditioning,* ii. *Current Research and Theory*. New York: Appleton-Century-Crofts, 64–99.

ROESE, N. J. (1997). 'Counterfactual Thinking', *Psychological Bulletin* 121: 133–48.

ROSER, M. E., FUGELSANG, J. A., DUNBAR, K. N., CORBALLIS, P. M., and GAZZANIGA, M. S. (2005). 'Dissociating Processes Supporting Causal Perception and Causal Inference in the Brain', *Neuropsychology* 19: 591–602.

ROZENBLIT, L., and KEIL, F. C. (2002). 'The Misunderstood Limits of Folk Science: An Illusion of Explanatory Depth', *Cognitive Science* 26: 521–62.

SAFFRAN, J. R., ASLIN, R. N., and NEWPORT, E. L. (1996). 'Statistical Learning by 8-Month-Old Infants', *Science* 274: 1926–8.

—— JOHNSON, E. K., ASLIN, R. N., and NEWPORT, E. L. (1999). 'Statistical Learning of Tone Sequences by Human Infants and Adults', *Cognition* 70: 27–52.

SATPUTE, A. B., FENKER, D. B., WALDMANN, M. R., TABIBNIA, G., HOLYOAK, K. J., and LIEBERMAN, M. D. (2005). 'An fMRI Study of Causal Judgments', *European Journal of Neuroscience* 22: 1233–8.

SCHLOTTMANN, A. (1999). 'Seeing It Happen and Knowing How It Works: How Children Understand the Relation between Perceptual Causality and Underlying Mechanism', *Developmental Psychology* 35: 303–17.

—— (2000). 'Is Perception of Causality Modular?' *Trends in Cognitive Sciences* 4: 441–2.

—— and ANDERSON, N. H. (1993). 'An Information Integration Approach to Phenomenal Causality', *Memory & Cognition* 21: 785–801.

—— and SHANKS, D. R. (1992). 'Evidence for a Distinction between Judged and Perceived Causality', *Quarterly Journal of Experimental Psychology* 44A: 321–42.

—— RAY, E. D., MITCHELL, A., and DEMETRIOU, N. (2006). 'Perceived Physical and Social Causality in Animated Motions: Spontaneous Reports and Ratings', *Acta*

Psychologica 123: 112–43.

Scholl, B. J., and Nakayama, K. (2002). 'Causal Capture: Contextual Effects on the Perception of Collision Events', *Psychological Science* 13: 493–8.

—— and Tremoulet, P. D. (2000). 'Perceptual Causality and Animacy', *Trends in Cognitive Sciences* 4: 299–309.

Schulz, L. E., and Gopnik, A. (2004). 'Causal Learning across Domains, *Developmental Psychology* 40: 162–76.

Shanks, D. R. (1995). 'Is Human Learning Rational?' *Quarterly Journal of Experimental Psychology* 48A: 257–79.

Sloman, S. A. (2005). *Causal Models: How People Think About the World and Its Alternatives*. Oxford: Oxford University Press.

—— and Hagmayer, Y. (2006). 'The Causal Psycho-Logic of Choice', *Trends in Cognitive Sciences* 10: 407–12.

Sobel, D. M., and Kushnir, T. (2006). 'The Importance of Decision Making in Causal Learning from Interventions', *Memory & Cognition* 34: 411–19.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, Prediction, and Search*. Berlin: Springer.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., and Blum, B. (2003). 'Inferring Causal Networks from Observations and Interventions', *Cognitive Science* 27: 453–89.

Tenenbaum, J. B., and Griffiths, T. L. (2001). 'Structure Learning in Human Causal Induction', in T. Leen, T. Deitterich, and V. Tresp (eds.), *Advances in Neural Information Processing Systems*. Cambridge, Mass.: MIT, xiii. 59–65.

Tomasello, M., and Call, J. (1997). *Primate Cognition*. Oxford: Oxford University Press.

Turner, D. C., Aitken, M. R. F., Shanks, D. R., Sahakian, B. J., Robbins, T. W., Schwarzbauer, C., and Fletcher, P. C. (2004). 'The Role of the Lateral Frontal Cortex in Causal Associative Learning: Exploring Preventative and Super-Learning', *Cerebral Cortex* 14: 872–80.

Tversky, A. (1977). 'Features of Similarity', *Psychological Review* 84: 327–52.

Uleman, J. S., Saribay, S. A., and Gonzalez, C. M. (2008). 'Spontaneous Inferences, Implicit Impressions, and Implicit Theories', *Annual Review of Psychology* 59: 329–60.

Van Hamme, L. J., and Wasserman, E. A. (1994). 'Cue Competition in Causality Judgments: The Role of Nonpresentation of Compound Stimulus Elements', *Learning and Motivation* 25: 127–51.

Waldmann, M. R. (1996). 'Knowledge-Based Causal Induction', in D. R. Shanks, K. J. Holyoak, and D. L. Medin (eds.), *Causal Learning: The Psychology of Learning and Motivation*. San Diego, Calif.: Academic Press, xxxiv. 47–88.

—— and Hagmayer, Y. (2006). 'Categories and Causality: The Neglected Direction', *Cognitive Psychology* 53: 27–58.

White, P. A. (1995). *The Understanding of Causation and the Production of Action*. Hillsdale, NJ: Erlbaum.

—— (2003*a*). 'Causal Judgement as Evaluation of Evidence: The Use of Confirmatory and Disconfirmatory Information', *Quarterly Journal of Experimental Psychology*

56A: 491–513.

—— (2003*b*). 'Effects of Wording and Stimulus Format on the Use of Contingency Information in Causal Judgment', *Memory & Cognition* 31: 231–42.

—— (2003*c*). 'Making Causal Judgments from the Proportion of Confirming Instances: The pCI Rule', *Journal of Experimental Psychology: Learning, Memory, & Cognition* 29: 710–27.

—— and MILNE, A. (1997). 'Phenomenal Causality: Impressions of Pulling in the Visual Perception of Objects in Motion', *American Journal of Psychology* 110: 573–602.

—— (1999). 'Impressions of Enforced Disintegration and Bursting in the Visual Perception of Collision Events', *Journal of Experimental Psychology: General* 128: 499–516.

WOLFF, P. (2007). 'Representing Causation', *Journal of Experimental Psychology: General* 136: 82–111.

CHAPTER 22
CAUSATION AND OBSERVATION

HELEN BEEBEE

1. INTRODUCTION

Hume argued (or so we are typically taught as undergraduates) that since no intrinsic causal relation can be observed, no idea of causation can be derived from experience, and so causation, conceived as an intrinsic relation between causes and effects, cannot exist—or, at the very least, we have no grounds for believing in its existence. Hence we must either do without the concept altogether, or else attempt to analyse causation in such a way that the concept reduces to notions that are empirically respectable. And, given that the only intrinsic relations that count as empirically respectable are spatio-temporal contiguity and temporal priority, our analysis of causation is going to have to appeal to an extrinsic relation, namely, the instantiation by the cause-and-effect pair of a regularity. (For discussion of the intrinsic/extrinsic distinction, see for example Langton and Lewis (1998) and Menzies (1999).)

Hume's (alleged) argument has had a central place in the history of the development of 'Humean' analyses of causation—analyses that construe causation as an extrinsic relation whose obtaining depends on patterns of regularity (see Ch. 7 above). It is therefore surprising that its central premiss—the claim that causation cannot be observed—has, until relatively recently, been the focus of very little philosophical attention. Instead, the claim was mostly simply accepted as obvious: 'As we have known from Hume,' Alvin Goldman says, 'causal connections between events cannot be directly observed' (Goldman 1993: 373). David Armstrong (1997: 211), quoting the same passage, notes that Goldman 'had the bad luck here to stand for whole generations of analytical philosophers'.

There were, of course, some philosophers who denied that causal relations are unobservable. C. J. Ducasse (1965: 177) says, '[t]he plain fact ... is that everyone has *perceived*—and I say *perceived*, not *inferred*—that, for example, a particular tree branch was *being caused to bend* by a particular bird alighting on it.' Elizabeth Anscombe is rather more well-known for making the same point. Hume 'confidently challenges us', Anscombe ([1971] 1993: 93) says, 'to "produce some instance, wherein the efficacy is plainly discoverable to the mind, and its operations obvious to our consciousness or sensation" [Hume [1739–40] 1978: 157–8]. Nothing is easier: is cutting, is drinking, is purring not "efficacy"?' Before both Ducasse and Anscombe, David Armstrong (1962: 23) had claimed that 'tactual perception ... gives us immediate awareness of objects *making things happen* to our body'. But he was not trying to make trouble for the Humean, explicitly saying, 'I do not think that this necessarily contradicts a Humean or semi-Humean analysis of the *nature* of causation'(ibid.).

Ducasse and Anscombe *were* trying to make trouble for the Humean, however. Anscombe rightly identified the claim that causal relations cannot be observed as a central plank in the standard argument for the regularity theory of causation, and in challenging that claim she challenged the standard objection to the kind of 'singularist' view that both she and Ducasse endorsed. (For the purposes of this chapter, 'singularism' is the view that causal relations are intrinsic relations, and so their obtaining does not depend on the relevant events instantiating a regularity.) Ducasse went even further: he explicitly used the allegedly obvious fact that we can observe causation to argue that the Humean view that there is no intrinsic causal relation between causes and effects is false, and singularism is true. 'Causation is therefore not to be confused with causal law, as too often is done,' he says. 'An empirically discovered causal law is causal not because it asserts a uniformity of sequence ..., but because it is an induction from perceived occurrences each of which, *in its own individual right,* was a case of causation and was perceived to be so' (Ducasse 1965: 178). Armstrong (1993: 170) later followed Ducasse's lead, writing that the regularity theory of causation 'runs counter to what seem to be the plain facts of experience. Just as we are directly aware of our own mental states, so, it seems to me, we are directly aware of certain cases of token causality. (We could not be directly aware of a cosmic regularity.)'

I shall argue in the course of this chapter that Anscombe and Ducasse may well be right, since the evidence from psychology suggests that we can indeed have experiences that represent the scene before our eyes as causal; and Anscombe was right to claim that this undercuts one traditional way of arguing for a Humean view. On the other hand I shall argue, *pace* Ducasse and Armstrong, that there is also no good argument from the *observability* of causal relations to their being intrinsic. In other words, the question of whether or not causation is observable turns out to be largely irrelevant to metaphysical issues concerning the nature and existence of causation.

I proceed thus: in sect. 2, I briefly discuss Hume's original argument concerning the absence of a sensory impression of causation. Hume's argument is important not just because of its historical significance in the debate about the observability of causation, but because it raises issues that arise within that debate in a particularly pure form. In sect. 3, I consider several ways in which psychologists and philosophers have attempted to characterize the sense in which causation might be 'observable', and the implications for the viability of a regularity account of causation. In sect. 4, I consider whether causation can be experienced in non-visual cases, specifically the experience of touch and the experience of agency. In sect. 5, I consider briefly whether the observability of causation makes trouble for broadly Humean, non-regularity accounts of causation, namely counterfactual, projectivist, and agency theories of causation.

2. HUME'S ARGUMENT

Hume's famous discussion of the origin of the idea of necessary connection ([1739–40] 1978: Bk. 1 pt. 3. §14 and [1748/51] 1975: §7) is often cited in discussions of the observation of causation, and it is undeniable that belief in his claim that 'all events seem entirely loose and separate' ([1748/51] 1975: 74) has acted as a strong motivational force for many Humeans. In this section I analyse Hume's argument and argue that, from a contemporary

perspective, it is not at all persuasive.

Hume's stated aim, in the sections 'Of the idea of necessary connexion' in the *Treatise* and first *Enquiry*, is to find the impression-source of our idea of necessary connection—the thought being that any *bona fide* idea must have its source in an impression, and so if we are legitimately to deploy the idea of necessary connection in our talk and thought, we must trace it back to its source. At the point where he undertakes this task in the *Treatise*, he takes himself already to have established that causes must be spatio-temporally contiguous with, temporally prior to, and constantly conjoined with their effects ([1739–40] 1978: 74–7). But clearly the idea of causation also includes the idea of necessary connection, and so the source of that idea must be traced to an impression of some kind.

Hume's argument runs roughly as follows.

(1) On first observing an event of a given kind, we cannot predict with certainty what effect will follow.

(2) If we could perceive a necessary connection between causally related events, then we *would*, on first observing the cause, be able to predict with certainty what effect would follow.

Therefore

(3) We cannot perceive a necessary connection between causally related events.

Therefore

(4) The impression-source of the idea of necessary connection (if there is such a source) cannot be some property or relation available to the senses.

Therefore

(5) Meaningful deployment of the idea of necessary connection—if it is possible at all— requires that we do not attempt to refer to any necessary connection, or power or force or efficacy, residing 'in the objects', since such usage could only be legitimate if the idea of necessary connection had as its source an impression of sensation (that is, an impression that genuinely detected necessary connections 'in the objects'). The only mind-independent features of the world that the concept of causation can successfully refer to are thus contiguity, temporal priority, and constant conjunction.

Hume's argument has thus been taken to be an argument against singularism: causation, if it exists at all, cannot be conceived as an intrinsic relation between causes and effects.

Three preliminary comments are in order. First, I have ignored some crucial elements in

Hume's overall argument. In particular, I have ignored his arguments that neither our control over our thoughts nor our control over our actions can provide the impression-source for the idea of necessary connection. (Hume's grounds here are analogous to those adduced for (3) above: if, for example, I could detect a genuine necessary connection between my wanting or deciding to raise my arm and my actually raising it, I would be able to tell just by experiencing my own act of will that my arm is in fact going to rise up. But I cannot do this; I might discover, for example, that my arm is paralysed and I am unable to raise it.)

Second, whether Hume really makes the final step in the argument is a controversial question. Galen Strawson and John Wright, for example, both argue that Hume holds that we *do* refer in our causal talk and thought to genuine, mind-independent, intrinsic causal relations, even though our *idea* of necessary connection is not derived from an impression of sensation (see Strawson 1989; Wright 2000; see also my 2006: ch. 7 for discussion).

Finally, Hume himself does not simply stop at (5). He goes on to claim that the idea of necessary connection *does* have an impression-source; it's just that the impression is an impression of 'reflection' rather than sensation. Roughly, the idea is that after we have observed cause-and-effect pairs of the same kind several times, we acquire the habit of inferring the effect from the cause. At the very same time, the impression of necessary connection appears: 'when one particular species of event has always, in all instances, been conjoined with another, we make no longer any scruple of foretelling one upon the appearance of the other, and of employing that reasoning, which can alone assure us of any matter of fact or existence. We then call the one object, *Cause*, the other, *Effect*' ([1748/51] 1975: 75). Thus, Hume claims, the impression of necessary connection just *is* the impression that arises when, on observing an event of kind *A* that we have previously seen to be immediately followed by an event of kind *B*, we infer or come to expect that an event of kind *B* will occur.

It is controversial whether Hume thinks that, having uncovered the internal impression-source for the idea of necessary connection, meaningful deployment of that idea is possible. On the one hand, he clearly does not have a subjectivist view of causation, which is what would seem to be required if the idea of necessary connection is an essential part of the idea of causation and the former idea *refers* to an inner mental state. This consideration (amongst others) might lead one to a regularity-theory interpretation of Hume, according to which causation really is just a matter of contiguity, temporal priority, and constant conjunction, and not, additionally, a matter of necessary connection. On the other hand, he does not appear to advocate revising the concept of causation in a way that excises the troublesome idea of necessary connection. These considerations might lead one in the direction of a sort of dispositionalist or secondary-quality interpretation, so that causation is defined as that feature of the world that generates the impression of necessary connection; or, alternatively, a projectivist interpretation, according to which causation is a projection of the idea of necessary connection onto a world of brute regularities (see Coventry 2006). (For discussion of these interpretative options, see Beebee 2006: chs. 5 and 6.)

Setting these issues aside, how does Hume's argument—as described above—fare from a contemporary perspective? Badly, it turns out, and for a number of reasons. Let's take the steps in the argument one at a time.

The first premise is hard to deny, and not merely because it is hard to see how a decisive counterexample might be found, given that we are exposed to so many regularities from the

moment we are born (see sect. 3.4 below). Hume's positive argument for (1) is that the existence of a cause and the absence of its effect is always conceivable. This shows that the cause does not entail the effect, and so the effect cannot be inferred a priori from the cause.

Several authors have pointed out that (2) is unwarranted, however. J. L. Mackie, for example, calls the kind of necessity that would ground a priori inference from causes to effects (that is, inference from cause to effect just on the basis of one-off observation of the cause) 'necessity$_2$'. (Necessity$_2$ would have to be some feature *of the cause*—a 'power of production', as Hume sometimes calls it—such that discerning that feature would enable us to figure out a priori what the cause would bring about.) But he points out that an intrinsic causal connection between causes and effects need not be of the kind that would license a priori inference; he calls a causal connection of this weaker kind 'necessity$_1$'. (Necessity$_1$ would be a genuine *relation* between cause and effect, rather than a power residing in the cause.) Mackie argues that (2) holds only for necessity$_2$: if we do not presuppose that causal necessity must license a priori inference, but accept instead that it could be a relation such as power or force or efficacy (all of which Hume himself takes to be synonymous with 'necessary connection'), then (2) is false (Mackie 1974: ch. 1; see also Menzies 1998: 344–5).

As an objection to Hume himself, Mackie's objection arguably fails. Hume takes it for granted that causation is what grounds *all* our empirical reasoning: when we draw conclusions about what will happen next, or what is going on in the next room, or what happened in 1066, we are reasoning from causes to effects, or vice versa. For Hume, no relation that failed to be essentially tied to inference could fill that role. To put it another way, Mackie is surely right that, granting (1), we might nonetheless be able to detect a *relation* between causes and effects, even if we cannot detect anything in the *cause itself* that licenses inference to the effect. But no such relation could provide the foundation Hume is looking for. A relation that we can observe to obtain only when we see *both* the cause, *a, and* the effect, *b*, cannot, just by itself, generate any inferences at all, since the next time we come across an *A*, we will have no grounds for supposing that it will similarly cause a *B*. In effect, the problem is that, while we might in principle be able to infer that event *a* instantiated *some* kind that is (observably) necessarily connected to events of some kind that *b* instantiated, we will not have any idea what the relevant kinds are; so the next time we come across an event similar in all or some observable respects to *a*, we will not have any idea whether the new event is of the kind that is necessarily connected to *B*s. Only sufficient experience of constant conjunction could inform us of what the relevant kinds are. But in that case, the inference from cause to effect would have to rely on 'habit' or 'custom'—that is, inference on the basis of observed regularity. So the detection of necessity$_1$ would play no essential role in inference from causes to effects, since habit or custom is perfectly capable of generating the required inference without the aid of such detection.

Mackie's objection is perfectly legitimate, however, when aimed at someone who thinks that Hume's argument establishes the falsity of singularism, which merely requires the existence of an intrinsic causal relation, and not the existence of something essentially tied to inference. If one wants to claim that a causal relation of *this* kind cannot be observed, then one cannot appeal to (2) in order to establish (3). In fact, many Humeans have taken it to be just obvious, presumably on independent, phenomenological grounds, that (3) is true. Whether

there really are such phenomenological grounds remains to be seen, of course, and I shall return to this question later; but it is worth pointing out that Hume himself does not have anything terribly convincing to say on the matter. He does say, earlier in the *Treatise*: 'Motion in one body is regarded upon impulse as the cause of motion in another. When we consider these objects with the utmost attention, we find *only* that one body approaches the other; and that the motion of it precedes that of the other, but without any sensible interval' ([1739–40] 1978: 76–7; my italics). But this is a phenomenological claim that the singularist can simply deny; and as we have seen, Ducasse and Anscombe do deny it. Indeed, one might speculate that Hume is sufficiently wedded to the claim about the foundational status of causation in empirical reasoning that it does not even occur to him to seriously consider, from a purely phenomenological point of view, whether the observable relations between causes and effects might not be restricted to temporal priority and contiguity.

The move from (3) to (4) is perhaps uncontroversial; unfortunately, however, the move from (4) to (5) is highly implausible by contemporary lights. For suppose we grant that there is no sensory impression of necessary connection (whether we read 'necessary connection' as the kind of feature that could in principle generate a priori inference (necessity$_2$) or simply as an intrinsic causal relation between causes and effects (necessity$_1$)). This claim only entails that we cannot meaningfully refer to necessary connections (in either sense) if we adopt a very strict meaning-empiricism, according to which an idea or concept simply lacks meaning if it is not grounded in a sensory impression. The view that the meaning of a term must be grounded in direct observational contact with the entity or relation referred to has long been out of favour, for very sensible reasons. (For discussion of this in relation to Hume's argument, see Menzies 1998: 356–9.) So even if Hume were right about (4), there would still be no grounds for believing (5).

All things considered, then, Hume's argument for the claim that we cannot meaningfully refer to necessary connections in nature is utterly unpersuasive from a contemporary perspective; and taken as an argument that we cannot even refer to some intrinsic causal relation between events that is weaker than necessity (such as production or bringing-about)—that is, as an argument against singularism—it is even worse.

Where does this leave the prospects for a Humean who claims support for her view from the unobservability of causation? The best that can be done, I think, would be to argue that there are independent phenomenological grounds for holding that causal relations cannot be observed, and then to attempt to argue that, while this does not provide any justification for holding that the claim that intrinsic causal relations exist is meaningless, it justifies the claim that we have no good grounds for *believing* in them. Unfortunately, however, whatever the prospects for the second stage of the argument, the prospects for the first stage—for arguing that causation cannot be observed—are not good. I examine the evidence for this in sect. 3 below.

It is worth noting, however, that there is some evidence that even Hume thought that we can have visual experiences as of one thing causing another. Most of us know that Hume said, 'all events seem entirely loose and separate' ([1748/51] 1975: 74). What is generally forgotten is that Hume says this (and similar things) in the context of talking about 'single instances of the operation of bodies' (ibid. 73). In other words, he says that this is how things appear to us in situations where the impression of necessary connection is not present because the habit of

inferring the effect from the cause has not yet been established. This leaves open the possibility that, for Hume, once the habit *has* arisen, and the corresponding impression of necessary connection is present, the phenomenology changes: events that previously seemed loose and separate no longer seem so. In other words, it is entirely possible to attribute to Hume the view that (once the inferential habit has been established) we do, in fact, have experiences as of one event causing another. As Blackburn puts it, we fail to engage with Hume 'if we merely insist, as many thinkers do, that we properly describe the *perceived* states of affairs in causal terms—*see* bricks splashing in water, balls breaking windows, things pushing and pulling' (Blackburn 1984: 211–12; see also Wright 2000; Kail 2001; Beebee 2006: ch. 4).

Of course, for Hume the impression does not *detect* genuine necessary connections: it is an impression of reflection, arising thanks to the inference we draw from cause to effect, and not an impression of any intrinsic feature of the scene before our eyes. But, once we abandon the claim that an impression that genuinely detected causation would have to license a priori inference from cause to effect, the only justification we can take from Hume for holding that there is no *sensory* impression of causation is the brute claim that, *on first observing them,* 'all events seem entirely loose and separate' (the thought being that since exactly the same outward features are present each time, a *sensory* impression of causation ought to be present on first observing a pair of events of a certain kind, rather than arising only once several instances have been observed). I return briefly to the question whether there is any evidence for the truth of this claim in sect. 3.4 below.

3. EXPERIENCE OF CAUSATION

It is a good idea to get a handle on the experience of causation that is independent of our metaphysical views about causation. Indeed, this is mandatory if we are to stand any chance at all of coming to any metaphysical conclusions based on claims about what we 'see' when we look at scenes that are, in fact, causal. For example, it is no good trying to argue that, since there is overwhelming psychological evidence that we perceive causation, and since one cannot perceive that *p* unless it is true that *p*, eliminativism about causation must be false. The eliminativist will legitimately respond that, whatever the psychological evidence may be, it cannot, just by itself, constitute evidence that there really *is* any causation. When psychologists talk about the *perception* of causation, they are not using 'perception' as a success term. (This is obvious from the fact that psychologists happily talk about perception even in cases of illusion.) I shall use 'perception' in this sense throughout, and take it to be synonymous with the notion of *causal experience.*

Imagine watching a game of snooker. A player lines up her shot, hits the white ball with the cue stick, the white ball moves towards the black ball, and makes contact with it; whereupon the black ball moves towards, and then into, the pocket. Call an experience that *just* represents the kinematic features of this scene—the positions and movements of the cue stick and the balls—a *thin* experience. Call an experience, or a belief, that additionally represents the scene as having causal elements—for example, that the impact of the cue stick *makes* the white ball move or *causes* it to move—a *causal* experience or belief. A standard (though, as we shall see,

mistaken) view amongst philosophers is that Humeans are required to deny that causal experience is possible: we have causal *beliefs*, but these beliefs must be *inferred* from thin experiences together with background beliefs—beliefs about regularities, for example.

The argument for this claim goes something like this: for a Humean—or at least for a regularity theorist (see Ch. 7 above)—the causal relation is extrinsic: its obtaining between two particular events *a* and *b* depends upon systematic patterns of regularity. But since the regularity itself is manifestly not present in the scene before one's eyes, the causal relation is no part of the visual stimulus on the basis of which one comes to say such things as 'I saw the white ball propel the black ball into the pocket' (as opposed to 'I saw the white ball touch the black ball, and then the black ball moved into the pocket'). Hence such a report must be inferred on the basis of the thin, kinematic experience together with antecedently held beliefs about regularities: it cannot itself be a report of a causal experience. Thus Armstrong writes:

Suppose ... that one starts ... from the *premiss* that singular causes are no more than instantiations of cosmic, or at least very widespread, uniformities. Then of course it *must* be the case that recognition of singular causation is a more or less sophisticated inference triggered off in us by the perception of the current sequence plus memory of the outcome of other such sequences. (1997: 213–14; see also Sosa and Tooley 1993: 13; Cartwright 2000)

The problem with this argument is that it trades on an undefined notion of 'perception'. Suppose that by 'perception of the current sequence' we take Armstrong to be referring to the visual stimulus that elicits causal judgements. Well, visual stimuli quite generally massively underdetermine the representational content of one's experiences (see Fodor 1984). If we define 'perception' in such a way that *only* the visual stimuli that we respond to count as candidates for being experienced, virtually everything we normally take ourselves to experience would turn out to be inferred rather than experienced. The claim that there are balls on the table, for example, would turn out to be inferred, since the visual stimulus at any given moment only includes whichever faces of the balls happen to be pointing in my direction. So, according to this conception of experience, I experience the facing surfaces of the balls, and infer from that, together with background beliefs, that there really are balls there.

Of course, the underdetermination of experience by visual stimuli might be taken to show that in a sufficiently broad sense of 'inferred', all causal content is inferred, in the sense that a good deal of cognitive processing is needed for us to get from the visual stimuli to the experience. But this is an unhelpful sense of 'inferred' to use in the debate between those who think that causation is perceived (in the psychologists' sense) and those who think it is not. Philosophers sometimes talk about whether causation can be the object of 'direct awareness' or whether it can be 'perceived directly'—by which they presumably mean whether the causal relation, like the surface of a snooker ball, can be a visual (or tactile) stimulus (see e.g. Menzies 1993: 202–3; Armstrong 1997: 214–15). This may be a legitimate question, but it is a question that cannot plausibly be answered in the absence of an agreed metaphysical story about the nature of causation. It is uncontroversial that, say, the facing surface of a snooker ball is a part of the visual stimulus that results in us seeing the ball while the back is not; but there is no way, independent of an account of the metaphysics of causation, to establish

whether the causal relation can be a part of the visual stimulus that results in us seeing a sequence as causal. The singularist may say that it is, while the Humean will disagree; but it is clear that appeals to phenomenology or psychology will not resolve the issue.

We might instead interpret Armstrong as holding that only *local* features of the sequence can be perceived or experienced; this is a broader notion of perception than the one just discussed, since a local feature (for example the back of a snooker ball) need not be part of the visual stimulus. If locality of a feature is a necessary condition for experience or perception, then Armstrong will of course be right that a regularity view of causation entails that causation cannot be perceived. But, thus interpreted, Armstrong's argument still cannot be deployed as an objection to a regularity view, and for a familiar reason: on this conception of 'perception', we cannot establish whether causation can be perceived without first establishing whether or not the conditions required for causation to obtain are local conditions. So there is no prospect of arguing from the claim that causation can be perceived to the conclusion that the conditions required for causation are all local, since the latter claim is presupposed by the former.

What is needed, then, is a criterion for 'observability' or 'perceivability' or 'causal experience' that will allow us to establish, independently of any prior metaphysical commitments, whether causation can be observed or perceived, or, in other words, whether causal experience is possible. In the next three subsections, I discuss three such criteria.

3.1 Causal Perception and 'Encapsulation'

Imagine watching an image on a screen of two coloured squares: a green one on the left, and a red one in the centre. The green square moves across the screen towards the red one and stops when they touch, whereupon the red square moves off in the same direction—to the right. (This is known by psychologists as a 'launching event'.)

How would you describe your experience? Albert Michotte, in his *The Perception of Causality* (1946), ran the experiment, and elicited people's responses. Michotte reported that 'the observers see object *A* bump into object *B*, and *send it off* (or '*launch*' it), *shove it forward, set it in motion, give it a push*. The impression is clear; it is the blow given by *A* which *makes B go*, which *produces B*'s movement' ([1946] 1963: 20).

Michotte ran a lot of variants of the experiment, for example with varying lengths of delay between the contact of the green square and the motion of the red square, and, instead of having the green square move until it touched the red one, having spatial gaps of various sizes between the place where the green square stops and the red square begins. He found that the number of observational reports that invoked causal concepts dropped off sharply when the spatial or temporal gap between the two movements got big enough. He also ran experiments involving what he called 'qualitative causality'; for example, in one experiment a green circle is next to a red circle. The green circle suddenly changes colour from green to yellow, whereupon the red circle changes from red to blue. He found that subjects did not describe such sequences in causal terms: the 'most frequent impression was one of a succession of independent events' (ibid. 243). (See also Ch. 21 sect. 2 above.)

Michotte took himself to have made two important discoveries: first, that we do perceive—

that is, have experiences as of—one thing causing another (or, in Michotte's terms, we have a 'causal impression'), but, second, that such experiences are fairly tightly circumscribed: they only arise in cases that have certain kinematic features. He also hypothesized that the mechanism that generates the causal impression is innate (see Saxe and Carey 2006).

Let's concentrate for now on the first claim. Does the fact that, when asked to describe what they see in the basic case described above, subjects typically respond by invoking causal concepts such as 'pushing', 'shoving', and 'making the ball go', establish that causal experiences are possible? (Michotte's claim here is, of course, similar to Anscombe's and Ducasse's claims that we can straightforwardly perceive cutting, a branch being caused to bend, and so on.) The consensus is that the possibility of causal experiences is not established by such reports just on their own. Susanna Siegel provides a nice example of the dangers of reading off the contents of experience from observational reports: one person might report that a table appears to be, say, 5 metres away, while another, in the same situation, might report that it appears to be 7 metres away. But it would be rash to conclude that how the table appears really differs between the two people, since it might easily be that the difference in the reports is due solely to a difference in how good the observers are at judging distances (see Siegel 2009: 523).

One way to approach the issue is to consider the extent to which subjects' reports are affected by background information. One would expect that if the reports are really 'inferred' from background information, then one would be able to change the contents of the reports by varying the background information. For example, imagine seeing a cartoon of Jerry running around with Tom hot on his heels. It might be that observers describe what they see differently, depending on the preceding portion of the cartoon: if this involved Jerry just managing to escape from Tom's clutches, then they might describe it as Tom chasing Jerry. But if it involved Jerry tying an invisible wire to a sleeping Tom and then to himself, then they might describe it as Jerry pulling Tom along. If so, this would arguably show that chasing and pulling are not really 'perceived' features of the cartoon, since which feature is reported varies with different background information.

This conception of how to decide whether, or to what extent, a given element is part of our perceptual experience trades on the claim that perception is *modular*, and that perception is 'informationally encapsulated': that is, perceptual experience is not fully penetrated by all the perceiver's background information. The Mueller–Lyer illusion is a standard example—the lines still *look* to be different lengths to most people, even when they know that they are in fact the same length—whereas in the Tom and Jerry case, whether or not Tom is reported as chasing Jerry depends upon what the observer's background beliefs are. (See Fodor 1984 for a discussion of modularity in the context of the theory-ladenness of observation.)

In the causal case, then (assuming that perception is indeed modular), the question is whether subjects' *causal* reports of experience—reports that invoke causal concepts such as 'pushing', 'making the ball go', and so on—are sensitive to all their background information. Michotte's own experiments provide prima facie evidence that such reports are not sensitive to all the subject's background information, since the subjects are fully aware that the squares moving across the screen are not really causally interacting at all; it's not as though they are looking at actual interactions between moving blocks of wood, say. So one can think of Michotte's experimental set-ups as a kind of causal illusion: the sequences *appear* causal to

the observers, even though they know that they are not.

There is considerable psychological evidence that causation is indeed perceived in this sense (e.g. Schlottmann and Shanks (1992) describe an experimental situation in which short-term associative learning does not affect causal experience). As Brian Scholl and Patrice Tremoulet (2000: 306) put it, the 'phenomena of perceptual causality are mandatory in the way that most visual illusions are: to the degree that the events are clearly perceived ... the causal ... nature of the resulting percepts is nearly irresistible. This reflects a type of encapsulation: despite the fact that observers know that the displays are not really causal ..., this knowledge does not appear to be taken into account by the mechanisms that construct the percepts.' (See also Schlottmann 2000.)

3.2 Causal Experience and Phenomenal Difference

Susanna Siegel (2009) argues for a second way of answering the question, whether causation can be represented in experiences. Siegel deploys what might be called a 'method of phenomenal contrast'. The general idea is as follows. First, find a particular case where the very same observable situation might plausibly be capable of producing different experiences. One example Siegel (2009: 526) gives is playing catch indoors. You fail to catch the ball, which lands in a plant pot, and, just afterwards, the lights go out. Of course, you don't *believe* that the landing of the ball caused the lights to go out; nonetheless, it may *seem* to you that it did. In other words, 'the successive events seem to be unified in experience in a way that is not merely temporal'. On the other hand, it may equally not seem that way: your 'visual experience represents the ball's trajectory and its landing, and your visual experience represents the lights going out, but so far as your visual experience is concerned, these events merely occur in quick succession' (ibid.). The thought is that both such experiences are possible. Second, if this phenomenological claim is true, what does it show? Well, the phenomenal contrast requires explanation. If—as Siegel argues—the best explanation is that one experience represents the situation causally and the other does not, then we have good grounds for thinking that causal experience is possible.

Siegel considers and rejects two alternative explanations: that the 'unity' in the first case is somehow not *causal* unity, and second, that there are two components to the 'experience': a sensory element (that has no causal content) and a cognitive element (such as 'a disposition to form a causal belief'), so that what is lacking in the second case is cognitive rather than sensory (2009: 527). This leaves Siegel's own hypothesis, that the first case really is a case of an experience—and not of the two-component variety—that represents the scene as causal, as the best candidate explanation of the phenomenal contrast between the two cases.

Siegel's method provides an alternative to the encapsulation criterion described in the previous section. For Siegel, the issue is not whether or to what extent our observational reports can vary with differences in background information; the central phenomenological claim is that there can be phenomenal difference, holding everything else fixed—and so without any difference in background information. Whether Siegel's method provides a *better* alternative is a matter for dispute. One reason for scepticism is that the claim that there are, in fact, cases of phenomenal contrast is an empirical one for which there is no evidence beyond

the fact that, allegedly, we are able to *imagine* that one might, for example, see the ball landing and the lights going out as causally related, or, alternatively, one might not—given exactly the same background information. But it is unclear whether what the reader is casually able to imagine correlates especially well with the empirical facts of the matter. So empirical support is required if Siegel's method is to get off the ground, since we need to be sure that the facts for which attribution of causal experience is claimed to be the best explanation really are facts.

3.3 Causal Perception and 'Categorical' Perception

Stephen Butterfill (2009) advances a third strategy for arguing that causation can be perceived, which draws on an analogy with the psychology of speech perception. There is evidence that speech perception is 'categorical'. In one experiment, Alvin Liberman and Ignatius Mattingly (1985) presented subjects with twelve sounds, evenly distributed along the spectrum of sounds from 'da', through 'ga', to 'ba'. The difference between each sound and its neighbour cannot, in most cases, be detected; however, subjects are able to discriminate at two significant points: where they hear the sound changing from 'da' to 'ga' and where they hear it changing from 'ga' to 'ba'. In other words, the sounds in themselves do not fall into three distinct categories: there is no more difference between the last member of the first perceived category and the first member of the second perceived category than there is between any two adjacent members of the same category. And yet subjects report the sounds as falling into these three distinct categories: they notice distinctive changes at two points as they proceed along the spectrum. Moreover, these reported boundaries match up with the 'intended phonic gestures': 'da', 'ga', and 'ba'. What explains this coincidence? Liberman and Mattingly suggest that the best explanation is that the objects of speech perception are not the sounds themselves, but the intended phonic gestures. In other words, roughly speaking, we do not hear mere sounds and then *interpret* them as elements of speech (say, when hearing someone utter the sentence, 'there's a *ba*nana'); rather, we hear them *as* elements of speech.

Butterfill suggests that Liberman and Mattingly's strategy can be applied to the case of causal perception too. Experiential reports in Michotte-type sequences also exhibit category boundaries; for example, when successively longer delays are introduced in between the first circle making contact with the second and the second beginning to move, where the length of delay is increased by the same, minute, amount each time, the point at which experience stops being reported in causal terms does not vary much between observers. In the speech case, there were no category boundaries between the sounds themselves, but only between the sounds qua intended phonic gestures. Similarly, in the causal case, there is no boundary between the sequences that observers report in causal terms and those they do not, if we just consider kinematic features of the sequences; the increase in delay between the last sequence reported as causal and the first sequence not so reported is just the same as the increases in delay between the other members of the sequence. Now, suppose that there is evidence that the category boundary matches up with the conditions actually required for causal interaction (or perhaps with the conditions under which causal interactions occur according to our naïve conceptions, as is suggested by White and Milne (1999))—something that Butterfill says it is

'reasonable to conclude' (2009: 419). This correspondence needs to be explained, and, as with the speech case, the best available explanation is that causal interactions really are perceived.

3.4 Causal Experience and the Consequences for Metaphysics

There is, of course, a lot more to be said about the relative merits of the three different kinds of criteria described above. However, my focus here is on the implications of the claim that causal experience is possible for the metaphysical dispute about the nature of causation, and not on precisely how we ought to draw the line between what is 'perceived' and what is 'inferred'. So let us grant that causal experience is possible, in each of the three senses described above. Is this bad news for the Humean?

I claim not. Recall the prima facie problem faced by the Humean: if causation is an extrinsic relation (a matter of the instantiation of a regularity, say), then, since the existence of a regularity is not something that can be perceived, causation cannot be perceived. But causation *can* be perceived; hence Humeanism must be false. If this argument is to have any bite, the possibility of causal experience, in any or all of the senses described above, must be shown to be incompatible with the extrinsicality of the causal relation.

I shall discuss this issue only in relation to the encapsulation criterion, since this is a commonly accepted criterion and is also the one that has been used in psychological testing of causal perception (the results of which strongly suggest that causation is indeed perceived in this sense). According to this criterion, something can be deemed the content of experience if how things look to observers (as described in their observational reports) is not penetrated by all their background information—as with the Mueller–Lyer lines, which look different lengths even to observers who know that they are in fact the same length. Thus the claim that causation is perceived in this sense is *not* the claim that causal experiences can be had in the absence of *any* background information whatever. Were we to raise the bar that high, very few things would be capable of being represented in experience. (Consider someone looking sad, for example, or a table looking square. Without prior information about what sad people typically look like, or what square things look like when observed from an oblique angle, we would not be able to experience someone as looking sad or a table as looking square.) In any case, the claim that causation can be perceived in the absence of any background information is impossible to test, since *nobody*—or at least nobody who is capable of being the subject of a psychological experiment—has no background information whatsoever that is relevant to causation. So the fact (if it is a fact) that background beliefs about regularities are required in order for observers to report their experience in causal terms does not undermine the claim that causation can be perceived.

It is worth noting in this regard that there is a method—the 'looking time experiment'—for testing the perceptual experiences of babies as young as 4 months (see Spelke 1985 for a full account). Many such experiments focus on Michotte-type sequences. For example, Oakes and Cohen (1990) provide evidence that infants as young as 10 months can, in some cases, discriminate causal from non-causal sequences, just as adults do, and hence evidence that (in Michotte's terms) they, like adults, have a 'causal impression'. Rebecca Saxe and Susan Carey (2006: 151) note that these experimental results have since been reproduced in 7-month-old

infants.

What does this show? Well, Saxe and Carey (2006) evaluate the evidence for claims about the source of causal representations, and in particular for Michotte's hypothesis that the mechanism that generates the causal impression is innate. One of their conclusions is that, on the innateness issue, the existing psychological literature delivers no decisive evidence one way or the other. After all, 'by the time experimentalists can find robust evidence of causal perception, infants have already had six months of experience observing causal interactions' (ibid. 163). In other words—as I said above—nobody capable of being the subject of a psychological experiment has no background information whatsoever; and the background information available even to 6-month-old babies includes plenty of regularities, including broadly Michotte-type sequences.

In any case, even if we could somehow show that causal experience is possible in the absence of any relevant experience of regularities, it is unclear why this would be incompatible with a broadly Humean account of the nature of causation. It might be that human beings have an innate capacity to differentiate between the kinds of kinematic sequence that are, in fact, typically causal (as in the standard Michotte-type launching events with no spatial gap or temporal delay between contact of one object on the other and the movement of the second) and those that are not (as in Michotte's cases with significant spatial gap or temporal delay). This would in no way compromise the claim that sequences of the first kind are causal in virtue of instantiating regularities, rather than in virtue of an intrinsic relation that can be the object of 'direct awareness'.

On the other hand, the possibility of causal experience in the absence of background information concerning regularities *would* refute Hume's empirical claim that the 'impression of necessary connection' only kicks in once a habit of expectation—derived from past experience of regularity—has been established. It would therefore undercut (what is normally claimed to be) Hume's argument for the claim that causation is not an intrinsic relation (or that we have no evidence that such an intrinsic relation exists), which relies on the premiss that we cannot experience causation in 'single instances of the operation of bodies' (see Fales 1990: 23–30). And of course the less controversial claim that causal experience is possible at all—whether or not it is possible in the absence of background information about regularities —by itself undercuts the standard argument for Humeanism with which I started this chapter: that, since causation is unobservable *simpliciter*, we have good grounds for rejecting the view that causal relations are intrinsic (though we already saw in sect. 2 that there are independent reasons for rejecting this argument in any case).

4. NON-VISUAL CAUSAL EXPERIENCE

In this section, I briefly discuss two other ways, aside from the visual, in which causal relations have been thought to be perceivable: the cases of agentive (sect. 4.1) and tactile (sect. 4.2) experiences. The case of agentive causation does not, so far as I can tell, raise significantly different issues for the metaphysics of causation than does the visual case, so I shall not discuss its implications for metaphysics, except to point out a connection between agentive experience and agency theories of causation in sect. 5. The case of tactile experience has been much less discussed, and so I shall spell spend a little time discussing its

ramifications.

4.1 Agentive Experience

One part of Hume's ([1748/51] 1975: 64–9) argument that we do not have an impression of necessary connection that licenses a priori inference from causes to effects concerns the operation of the will. Predictably, the argument focuses on various things we would be able to know, were we to be able to discern a necessary connection between acts of will and (for example) bodily movements, which manifestly we do not know. I shall not describe Hume's argument or what is wrong with it here (but see Menzies 1998: 345–8 for a quick summary of both); what I am interested in is whether, by contemporary lights, there are any grounds for supposing that we *do*, in fact, have what (following Bayne 2008) I shall call *agentive experiences*: experiences as of acting intentionally, that is, as of bringing something about specifically *as an agent*.

Of course, if visual causal experiences of causation are possible, then plausibly we can at least sometimes visually experience our own bringing things about, as when I observe myself reaching out for a coffee cup and picking it up, say. But I shall not class such experiences as agentive experiences, unless the experience somehow also includes agency as such, as opposed to mere causation. That is, a visual experience that is *merely* an experience as of my hand reaching out and picking up the cup—something that is phenomenologically similar to that of watching a robot's hand doing the same—will not count as agentive. An experience will only count as agentive if its content includes my *performing an intentional action*.

An immediate problem faced by any attempt to argue that we do have distinctive agentive experiences is that in our everyday descriptions, we seem to be just as happy to describe other people's bodily movements in intentional terms as we do our own. If I am watching an auction and I see someone put a bid in by raising their arm, the most natural way to describe what I see is to say that *they raised their arm*, and not merely that their arm rose up. Indeed, it is easy to imagine that we will do this, in the right context, even if we know that the 'person' is really a remote-controlled robot. Whether or not we do indeed have experiences as of other people's acting intentionally is a question I shall leave open. The prima facie problem here, however, is that we are looking for a distinctively *first-personal* experience of agency: something that arises from *our own* 'acts of will'. So, if we are just as happy to describe other people's bodily movements in intentional terms as we do our own, we need to be careful not to rely solely on subjects' reports, since these may not discriminate between observation of others' intentional action and genuine, first-personal, agentive experience.

One experiment designed to provide evidence for the existence of agentive experience is Daniel Wegner's 'helping hands' experiment (Wegner 2002). Here is Siegel's (2005: 267) description of the experiment:

[P]erson *A* stands facing a mirror with arms inside a sleeveless robe, while person *B*, standing behind *A*, puts engloved arms through the arm holes so that *B*'s arms are where *A*'s would normally be. *B* then hears instructions directing the hands (e.g. to clap, wave, make a fist). People in *A*'s position who hear the directions report feeling a greater degree of control

o f *B*'s hands (3 on a 7-point scale) than do people in *A*'s position who do not hear the instructions (1 on a 7-point scale).

As Siegel notes, person *A* does not *believe* that she really is controlling *B*'s hands; hence 'this feeling is a candidate for being an experiential representation of efficacy' (ibid.) ('experience of efficacy' being Siegel's name for what I am calling 'agentive experience'). This connects with the modularity thesis described in sect. 3.1 above: the thought is that since how things *seem* is not penetrated by all the information in the subject's possession—she knows she is not in control but still has the *feeling* of control—the representation ('I am controlling *B*'s hands') would seem to be delivered by the perceptual system.

Tim Bayne speculates that dissociation between judgement and perception in the case of agency is also possible in two kinds of disorder: anarchic hand syndrome and utilization syndrome. He notes:

The two syndromes are similar in that each involves an inability to inhibit stimulus-driven actions. The patient with an anarchic hand ... will take food from another's plate ... ; the patient with utilization syndrome will put on multiple pairs of sunglasses, even when she is already wearing sunglasses. These actions ... may even be at *odds* with the patient's goals. (The patient doesn't want to take food from his neighbour's plate.) But despite their behavioural commonalities, these two disorders give rise to very different reports: whereas patients with utilization behaviour show no inclination to disown their actions, patients with an anarchic hand typically describe the hand as 'having a will of its own'. (Bayne 2008: 186)

Bayne contends that the differing judgements as to authorship of the actions are plausibly due to differences in phenomenology: the patient with an anarchic hand 'fails to experience himself as the agent of the movements of his anarchic hand', while the patient with utilization behaviour has normal experiences of agency. And the final step in the dissociation of perception and judgement would be to convince the patient with an anarchic hand that 'he is acting even though he doesn't experience himself as acting', or to convince the patient with utilization behaviour that 'she is not acting despite the fact that she experiences herself as acting' (ibid.). This would provide further evidence of encapsulation in the case of agency: evidence, that is, of the perceptual system delivering a verdict that is insensitive to what the patient believes to be the case.

Siegel herself (2005) deploys the same basic argument structure as she uses for the claim that we can perceive causation to argue that we can perceive what she calls our own 'efficacy'. Again, she gives a pair of examples where the stimuli and background knowledge remain the same, but where the phenomenology differs; she then offers the claim that one's own efficacy is experienced in one of the cases but not the other as the best explanation of the difference in phenomenology.

4.2 Tactile Experience

Rather less attention has been paid to the tactile case than to the visual and agentive cases. Nonetheless, both Evan Fales (1990: ch. 1) and David Armstrong (1997: 212–14) have taken the case of the sensation of pressure or force to strike a considerable blow against the Humean.

Fales (1990: 16) considers the case where someone pushes steadily against your forehead with their hand, and claims to identify several components of the 'intrinsic character of the sensation of force': (1) it has a spatial location; (2) it has a magnitude; (3) it has a direction in space; (4) several different forces can sometimes be differentiated; (5) felt forces form 'an algebra': 'they can be felt to add together in a certain way which depends upon their respective magnitudes and directions' (ibid.); and (6) asymmetry: '[t]hat production is an asymmetric relation is something we experience. We do not merely experience forces as having location, magnitude, and direction. We experience them as acting upon something ... ' For example, we 'are able to distinguish in perception between active agency on our part and the passive reception of force', and 'between an impressed force and the resistance of our bodies' (ibid. 17).

Fales proceeds to mount an argument against Hume's contention that we cannot, on first observing an event of a given kind, predict what will happen next, by asking us to consider whether a subject would be able to predict whether or in what direction his head would move if subjected to a blow to the forehead (see ibid. 23–30). I shall ignore this part of his argument, however, and concentrate on two questions. First, are Fales's phenomenological claims true? And second, if they are, do they pose any problems for the Humean?

We are by now familiar with the point that the claim that our experiences can have causal content does not serve to refute Humeanism. So in a sense it does not much matter, for current purposes, whether Fales is right about the phenomenology if we think of the phenomenological claims as being claims about the contents of tactile experiences.

On the other hand, the case of force or pressure might seem more worrying, because it might seem more obvious in this case that force or pressure is an object of 'direct awareness' rather than *merely* the content of experience, broadly conceived; and Humeans (conceived as those who take causation to be an extrinsic relation) must, of course, deny this. But in fact it is not really obvious at all that pressure is an object of direct awareness. Of course, it feels different when someone presses hard on your forehead to how it feels when they press more gently; and it feels different when they press on the left side of your forehead to how it feels when they press on the right side. But none of this establishes that the *pressing*—or its spatial location or its magnitude—is an object of direct awareness. After all, a flute sounds different when played centimetres away from your left ear to how it sounds when played in the next room; but this does not establish that the proximity of the playing is an object of direct awareness—though of course it might well be something that is *represented* in one's auditory experience. Moreover, in general—to reiterate a point made in sect. 3 above—it is unclear how one might establish what does and does not count as an object of direct awareness, in the absence of agreement about what there is in one's vicinity that is available to be a candidate for direct awareness; and such agreement is, of course, precisely what is lacking in the case of causation.

If Fales's argument is to have any impact on the metaphysical debate, then, it needs to be shown that force or pressure really is an object of direct awareness. Menzies (1993: 201–2) argues directly against this claim, saying that:

There is a counterfactual element to [causal] relations that cannot plausibly be claimed to be an object of direct awareness. Compare, for instance, the situation in which my sensation of pressure is caused by the impact on my body with the situation in which the sensation is actually caused by some other causal factor, coincidentally operating at the same time. These situations seem to differ only in terms of what is counterfactually true of them. The first situation is one in which it is true that if the bodily impact had not occurred, I would not have experienced the sensation of pressure, whereas the second situation is one in which this counterfactual is false. In determining the cause of my experience of pressure, I have to be able to determine whether this counterfactual is true or false. But it is clear that I cannot do this on the basis of my perceptual experiences, since the content of my experiences would be the same in both causal situations. It would seem, then, that ... I do not, after all, have direct, noninferential awareness of causation in this case.

Let's assume that by 'perceptual experiences' here, Menzies means 'direct awareness'. (I discuss whether counterfactual dependence can be experienced in a broader sense of 'experience' in sect. 5 below.) Armstrong (1993: 228) argues that Menzies' argument fails, because it could be that 'all that we are aware of is our body being pressed upon. Being pressed upon entails that something presses, so if our perception is veridical then something is doing the pressing. But perhaps, once collateral information (from other senses, etc.) is abstracted from, the pressure sensation involves nothing more than a quite indeterminate awareness of something or other doing the pressing.' It is unclear that Armstrong really addresses Menzies' concern here, however. Menzies' objection appears to be that the very same sensation can be produced either by genuine pressure—something pressing on me—or by something entirely different (some sort of direct stimulation of my brain by a neuroscientist, say), but where there is still, coincidentally, something pressing on my body. Genuine detection of pressure—*qua* causal relation—would require that I can tell the difference between these two cases, which I cannot do because the only relevant difference is a difference in the truth of the relevant counterfactual. Armstrong's suggestion appears to be that the argument fails because I cannot discern what, exactly, it is that is doing the pressing (I am only aware of 'something or other doing the pressing'); but Menzies' second case is supposed to be one where *nothing* that I have any awareness of is doing any pressing; the sensation of pressure is caused by something else. So pressing is not something I am directly aware of.

Menzies' argument fails, however, for another reason. It is no objection to the claim that we are directly aware of some feature F to say that a qualitatively identical experience as of F can be had in the absence of F itself. Consider the visual case of looking at a snooker ball: I can be directly aware of the part of the ball facing me (in the sense of the face being part of the visual stimulus). But this is perfectly consistent with there being possible cases where I have a qualitatively identical experience that is not caused by any small, red, hemispherical shape;

this is something I can perfectly easily hallucinate or dream, or perhaps have induced by my brain being tweaked by a neuroscientist. Of course, in the hallucinatory case I am not directly aware of anything at all; but this does not show that in the non-hallucinatory case I am not directly aware of anything either. Similarly for the case of pressure: the fact that I can have a tactile experience that is indistinguishable from genuine direct awareness of pressure, even though that experience is not *caused* by pressure, does not undermine the claim of the veridical case to be a genuine case of direct awareness.

This is not to say, however, that Fales has successfully shown that force or pressure *can* be an object of direct awareness. Indeed, as I have said, it is hard to see *how* this can be shown. Brute appeal to introspection is not good enough, for we have very good reasons to doubt the reliability of introspection in such matters. Consider, for example, the case of colours. Arguably, colours *seem* to us to be objects of direct awareness: intrinsic, categorical properties of objects whose nature is immediately given to us in experience. But a combination of scientific discovery and philosophical reflection have shown that this is a very difficult view to maintain: plausible candidates for being intrinsic colour properties of objects (surface reflectances, say) are equally *im*plausible candidates for being properties of which we are directly aware in sensory experience.

Now, Fales argues, in effect, that awareness of forces is not like this; on the contrary, the felt properties of forces listed above (or at least the first five of them) are precisely properties of forces as described by physics: 'they are exactly those which have been taken over into physics and given a precise representation there by means of a vector calculus' (Fales 1990: 16). But it is debatable whether the felt properties of forces really are as Fales describes them. For example, do felt forces really have a direction? If someone presses the end of my nose horizontally, it feels different to when they press the same point in a slightly upward direction. But this establishes nothing, for of course, I also have awareness of the movement of my nose as a result of the finger pressing on it, and it could just as well be that what I am *directly* aware of is simply the finger and the movement of my nose.

At any rate, the colour case shows that introspection is an unreliable guide to what we are and are not directly aware of, and that is enough to cast doubt on the claim that we are directly aware of forces. In the absence of a good argument that force or pressure is an object of direct awareness—as opposed to something that features in the *content* of experience, where (like the experience as of a billiard ball, as opposed to just its facing surface) the content of experience can be affected by background information—tactile experience of pressure, if it exists, is just another example of causal experience, and provides no more ammunition against Humeanism than do the visual and agentive cases.

5. OTHER BROADLY HUMEAN VIEWS

I have so far been characterizing 'Humean' views as, essentially, versions of the regularity theory of causation. But of course many philosophers who subscribe to a broadly Humean metaphysics do not hold a regularity theory of causation. Rival candidates include, in particular, counterfactual, projectivist and agency theories, and it is worth briefly discussing the connection between these views and the issues concerning causal experience.

Counterfactual theories of causation claim that causation is to be analysed, somehow or other, in terms of the notion of counterfactual dependence (see Ch. 8 above). We have already seen that Menzies claims that counterfactual dependence cannot be an object of direct awareness; does the same point apply to experience more broadly conceived? That is, can we not represent in experience something being such that, had it not happened, something else would not have happened either? At first sight, the natural answer is no; as Colin McGinn puts it (in the context of colour perception): 'Your eyes do not respond to *woulds* or *might have beens*' (McGinn 1996: 540). But recall that we are here not restricting the content of experience to 'what your eyes respond to'.

Siegel speculates that counterfactual dependence is sometimes capable of being represented in experience; for example, imagine a rock balanced on the tip of another rock. It is not obvious, Siegel thinks, that we cannot represent that scene as being such that the rock would tip over if pushed (2009: 532–3). But—as she notes—even if this is right, it only establishes that counterfactual experience is possible in cases where the antecedent ('if the rock had been pushed ...') is a 'natural continuation' of something that you see: you don't need to, as it were, imaginatively think away what is, in fact, in front of you. But many cases of counterfactual dependence are not like that; the possible situation that makes it true that, had I not hit the cue ball, it would not have struck the black, is one in which very little of the actual scene, where I do hit the cue ball, it moves across the table, and strikes the black, remains intact.

Psychologists have paid a good deal of attention to counterfactuals, but they have certainly regarded counterfactuals as falling on the side of thought rather than perception. So there is scant evidence one way or another on whether (a limited class of) counterfactuals might be capable of being represented in experience. This is, of course, an empirical question, but it seems unlikely that it has a positive answer. Just from everyday life it seems that 'observational' reports of counterfactual dependence are, at best, extremely rare. If this is right, then we cannot even make it to the starting line when it comes to asking whether such reports are genuinely reports of experience or not (as judged by, for example, whether they are affected by variations in background information), since there are no such reports to ask about in the first place.

From the point of view of the metaphysics of causation, though, does any of this matter? Suppose that we agreed that causal experience is possible, but counterfactual experience is not. Would this make trouble for a counterfactual analysis of causation? Arguably not. I have already argued that the fact that we cannot experience a sequence instantiating a regularity does not undermine a regularity theory of causation when combined with the fact that we *can* experience a sequence being causal; and the same basic point applies to the counterfactual analysis. Indeed, to put the point more vividly, if in general we required as a condition on an acceptable conceptual analysis C of some phenomenon X, where X is capable of being represented in experience, that C is also capable of being so represented, then conceptual analysis in general would be in big trouble. Consider my experience of seeing a person in front of me (assuming that such experience is possible). The kinds of psychological features that are standardly viewed as requirements on personhood are not, it seems, capable of cropping up in the content of experience. So—according to the line of argument under discussion— personhood cannot be conceptually analysed in psychological terms.

One might argue that there are good philosophical grounds for upholding the principle just

described; for example, one might subscribe to the view that the representational content of an experience is given by its truth conditions. But this would be to depart from the conception of 'perceivability' that generates a violation of the principle in the first place, for it would place constraints on what does or does not count as perceivable that are orthogonal to the constraints imposed by the kinds of psychological conceptions of perceivability that I have been discussing. Grant, for example, that observers' reports about causation are, but their reports about counterfactual dependence are not, informationally encapsulated. This is a purely psychological fact about the way our perceptual system works. To hold that the representational content of an experience is given by its truth conditions would be to place an additional constraint, beyond informational encapsulation, on what can and cannot be perceived; and so we would now have no reason to accept that the facts about informational encapsulation really reveal anything about what can and cannot be perceived. So we would have no reason to think that the above principle had really been violated.

A second broadly Humean view about causation—one that has a much more direct relation to the question about the possibility and nature of causal experience—is a projectivist view. On one interpretation of Hume, for example, Hume conceives of causation as a projection of our habits of expectation onto a world of loose and separate events; and that projection modifies our experience in such a way that events *seem* causally connected. Since this 'impression of necessary connection' provides the content for the *idea* of necessary connection, our causal experience plays an essential role in determining the meaning of our causal claims (see Beebee 2006: ch. 6)—though on such a view causation is not genuinely *perceived* in at least one sense, since what gives sequences their causal character is contributed by the mind of the observer rather than a detectable feature of the sequence itself.

A different version of projectivism (at least in a broad sense of 'projectivism') is the 'agency' view of causation defended by Peter Menzies and Huw Price (1993). Menzies and Price argue that causation should be seen as a secondary quality, analogous in some respects to colours, where the relevant experience, in which the concept of causation has its origin, is the experience of agency: of doing something as a means to achieving an end. They say:

[W]e all have direct personal experience of doing one thing and thence achieving 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 that does not depend on prior acquisition of any causal notion. (ibid. 194–5)

Arif Ahmed takes issue with this claim about the experience of agency:

I might agree with [the first sentence of the previous quote] if you take away two letters. What we all have direct experience of is doing one thing and *then* achieving another. I cannot see that we have direct experience of anything that distinguishes 'thence' with its causal implication from 'then' which lacks them. I cannot see that the sequences in which ends are brought about by means look any different from sequences in which the former merely *succeed* the latter. (Ahmed 2007: 125–6)

The discussions of Siegel (2005) and Bayne (2008) briefly described earlier bear, of course, on Ahmed's contention; they at least ought to make us consider whether it is as obvious as he claims that there is no such thing as the experience of agency. On the other hand, Menzies and Price seem to want more than mere experience of agency: they want 'direct non-linguistic acquaintance with the concept of bringing about'. This would appear to be a much stronger requirement, and arguably one that cannot be shown to obtain by the kinds of psychological study envisaged by Siegel and Bayne. This is because the notion of 'direct acquaintance' is, as we have seen, a metaphysically loaded notion: we could not, for example, be directly acquainted with bringing about, if bringing-about was merely a matter of the instantiation of a regularity, say. Siegel and Bayne are (explicitly or implicitly) concerned with a broader conception of experience or perception, of the kind whose content could in principle be established independently of metaphysical presuppositions.

That said, Ahmed argues that Menzies and Price's appeal to the experience of agency, as described above, is in fact an unnecessary hostage to fortune given their overall account, which trades on a conception of agency that is analysed as decision-making on the basis of what they call 'agent probabilities' (that is, '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' (Menzies and Price 1993: 190)). Ahmed (2007: 131) argues that, since someone could 'make judgements of agent-probabilities *without* being able to form causal judgements', agency, thus understood, can be thought of as conceptually prior, in the sense required for an agency theory of causation, to causation. If Ahmed is right, then an agency theory of causation does not require commitment to disputable claims about the experience of agency, although the agent's point of view—the perspective that delivers agent probabilities—remains a crucial part of the story (see Price 2007: esp. 279–83).

FURTHER READING

Hume's *Treatise* ([1739–40] 1978: Bk. 1 pt. 3 §14), and *Enquiry* ([1748/51] 1975: §7), are the starting points for the debate about causal experience and its connection to Humean metaphysics. Ducasse (1965), Anscombe ([1971] 1993), and Fales (1990: ch. 1) are attempts to tie causal experience to singularism; see also Menzies (1998), which includes a summary of some of the psychological literature on causal perception. Beebee (2003) argues that the evidence concerning causal experience provides no justification for either Humeanism or singularism. Bayne (2008) provides a good introduction to the issues concerning agentive experience, with plenty of references. Other relevant recent articles include Siegel (2009) and Butterfill (2009).

REFERENCES

AHMED, A. (2007). 'Agency and Causation', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Clarendon, 120–55.

ANSCOMBE, G. E. M. ([1971] 1993). *Causality and Determination*. Cambridge: Cambridge University Press; repr. in E. Sosa and M. Tooley (eds.), *Causation*. Oxford: Oxford University Press, 1993.

ARMSTRONG, D. M. (1962). *Bodily Sensations*. London: Routledge & Kegan Paul.

—— (1993). 'Reply to Menzies', in J. Bacon, K. Campbell, and L. Reinhardt (eds.), *Ontology, Causality and Mind: Essays in Honour of David Armstrong*. Cambridge: Cambridge University Press, 225–32.

—— (1997). *A World of States of Affairs*. Cambridge: Cambridge University Press.

BAYNE, T. (2008). 'The Phenomenology of Agency', *Philosophical Compass* 3: 182–202.

BEEBEE, H. (2003). 'Seeing Causing', *Proceedings of the Aristotelian Society* 103: 257–80.

—— (2006). *Hume on Causation*. London: Routledge.

BLACKBURN, S. (1984). *Spreading the Word*. Oxford: Oxford University Press.

BUTTERLL, S. (2009). 'Seeing Causings and Hearing Gestures', *Philosophical Quarterly* 59: 405–28.

CARTWRIGHT, N. (2000). 'An Empiricist Defence of Singular Causes', in R. Teichmann (ed.), *Logic, Cause and Action*. Cambridge: Cambridge University Press.

COVENTRY, A. (2006). *Hume's Theory of Causation: A Quasi-realist Interpretation*. London: Continuum.

DUCASSE, C. J. (1965). 'Causation: Perceivable? Or Only Inferred?', *Philosophy and Phenomenological Research* 26: 173–9.

FALES, E. (1990). *Causation and Universals*. London: Routledge.

FODOR, J. (1984). 'Observation Reconsidered', *Philosophy of Science* 51: 23–43.

GOLDMAN, A. I. (1993). 'The Psychology of Folk Psychology', in A. I. Goldman (ed.), *Readings in Philosophy and Cognitive Science*. Cambridge, Mass.: MIT, 347–80.

HUME, D. ([1739–40] 1978). *A Treatise of Human Nature*, ed. L. A. Selby-Bigge and P. H. Nidditch. 2nd edn. Oxford: Clarendon.

—— ([1748/51] 1975). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, ed. L. A. Selby-Bigge and P. H. Nidditch. 3rd edn. Oxford: Clarendon.

KAIL, P. J. E. (2001). 'Projection and Necessity in Hume', *European Journal of Philosophy* 9: 24–54.

LANGTON, R., and LEWIS, D. K. (1998). 'Defining Intrinsic', *Philosophy and Phenomenological Research* 58: 333–45.

LIBERMAN, A. M., and MATTINGLY, I. G. (1985). 'The Motor Theory of Speech Perception Revised', *Cognition* 21: 1–36.

MCGINN, C. (1996). 'Another Look at Color', *Journal of Philosophy* 93: 537–53.

MACKIE, J. L. (1974). *The Cement of the Universe*. London: Oxford University Press.

MENZIES, P. (1993). 'Laws of Nature, Modality and Humean Supervenience', in J. Bacon, K. Campbell, and L. Reinhardt (eds.), *Ontology, Causality and Mind: Essays in Honour of David Armstrong*. Cambridge: Cambridge University Press, 195–225.

—— (1998). 'Are Humean Doubts about Singular Causation Justified?', *Communication and Cognition* 31: 339–64.

—— (1999). 'Intrinsic versus Extrinsic Conceptions of Causation', in H. Sankey (ed.),

Causation and the Laws of Nature. Dordrecht: Kluwer, 313–29.

—— and H. PRICE (1993). 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science* 44: 187–203.

MICHOTTE, A. ([1946] 1963). *La Perception de la causalité*. Louvain: Institut Supérieur de Philosophie. English trans., *The Perception of Causality*, trans. T. R. Miles and E. Miles. Aylesbury: Hazell Watson & Viney, 1963.

OAKES, L., and COHEN, L. (1990). 'Infant Perception of a Causal Event', *Cognitive Development* 5: 193–207.

PRICE, H. (2007). 'Causal Perspectivalism', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Clarendon, 250–92.

SAXE, R., and CAREY, S. (2006). 'The Perception of Causality in Infancy', *Acta Psychologica* 123: 144–65.

SCHLOTTMANN, A. (2000). 'Is Perception of Causality Modular?', *Trends in Cognitive Science* 4: 441–2.

—— and SHANKS, D. (1992). 'Evidence for a Distinction between Judged and Perceived Causality', *Quarterly Journal of Experimental Psychology* 44: 321–42.

SCHOLL, B. J., and TREMOULET, P. D. (2000). 'Perceptual Causality and Animacy', *Trends in Cognitive Science* 4: 299–309.

SIEGEL, S. (2005). 'The Phenomenology of Efficacy', *Philosophical Topics* 33: 265–84.

—— (2009). 'The Visual Experience of Causation', *Philosophical Quarterly* 59: 519–40.

SOSA, E., and TOOLEY, M. (1993). 'Introduction', in Sosa and Tooley (eds.), *Causation*. Oxford: Oxford University Press, 1–32.

SPELKE, E. S. (1985). 'Preferential Looking Methods as Tools for the Study of Cognition in Infancy', in G. Gottlieb and N. Krasnegor (eds.), *Measurement of Audition and Vision in the First Year of Postnatal Life*. Norwood, NJ: Ablex, 323–63.

STRAWSON, G. (1989). *The Secret Connexion*. Oxford: Oxford University Press.

WEGNER, D. (2002). *The Illusion of Conscious Will*. Cambridge, Mass.: MIT.

WHITE, P., and Milne, E. (1999). 'Impressions of Enforced Disintegration and Bursting in the Visual Perception of Collision Events', *Journal of Experimental Psychology: General* 128: 499–516.

WRIGHT, J. P. (2000). 'Hume's Causal Realism', in R. Read and K. Richman (eds.), *The New Hume Debate*. London: Routledge, 88–99.

CHAPTER 23
CAUSATION AND STATISTICAL INFERENCE

CLARK GLYMOUR

1. INTRODUCTION

In the applied statistical literature, causal relations are often described equivocally or euphemistically as 'risk factors', or as part of 'dimension reduction'. The statistical literature also tends to speak of 'statistical models' rather than of causal explanations, and to say that parameters of a model are 'interpretable', often means that the parameters make sense as measures of causal influence. These ellipses are due in part to the use of statistical formalisms for which a causal interpretation is wanted but unavailable or unfamiliar, and in part to a philosophical distrust of attributions of causation outside experimental contexts, misgivings traceable to the disciplinary institutionalization of claims of influential statisticians, notably Karl Pearson and Ronald Fisher. More candid treatments of causal relations have recently emerged in the theoretical statistical literature.

2. CAUSAL INTERPRETATIONS OF STATISTICAL MODELS

In statistics, differences in the family of probability distributions considered are almost always accompanied by a new 'model' terminology, with the result that similarities and dissimilarities relevant to causal inference and prediction are sometimes obscured. Some of the most common frameworks include:

1 . *ANOVA models*. Analysis of Variance is one of the most widely used methods for estimating the effect of one or more categorical variables on a continuous variable. A unit u_{ij} belongs to some group j having the same value for the potential cause, *X*, and the value of *Y* for u_{ij} is Y_{ij}. The model is

(1) $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$

where μ is the average value across all values of *X*, α_j is the mean of the group with the jth value of *X*, and $\in_{ij}$ is the value of a Normally distributed random variable, with the same distribution for all units. The intuitive causal interpretation, where appropriate (which is not always), is that moving units from one group to group j, or intervening to give new units the jth value of *X*, would on average result in values of *Y* characterized by α_j.

2. *Recursive, linear structural equation models*. Variables are ordered and each is written as a linear function of a subset of its non-descendants in the ordering plus 'noise'. Noises may be independently distributed or correlated. Variations include treating values of linear coefficients as random variables, and the use of binary variables as sources (exogenous variables). Some variables may be unrecorded or 'latent'. The causal interpretation is that the equations (with fixed coefficients) specify the change that would occur in the dependent variable Y for a forced, exogenous unit change in any independent variable X occurring in the equation for Y, if other variables (including the noise variable for Y) were held constant by intervention. With random coefficients estimated by their means, the equations give the average change in Y for a unit forced, exogenous change in X (with similar constraints on other variables except the noise term). This class of statistical models includes as special cases factor analysis models, linear regression models, and principal components models.

3. *Non-recursive linear structural equation models* do not require an ordering of variables: X may occur in an equation for Y, and Y may reciprocally occur in an equation for X, and more generally a chain of equations may occur constituting (in graphical terms) a closed path from X to X. One causal interpretation is that each variable X corresponds to a time series X_t and if an equation such as $Y = a X + \in$ occurs, then $Y_t = aX_t + \in_t$ in the time series, with the $\in$ variables all independently distributed. An intervention on X then fixes (or randomizes) X_t at some arbitrary time t_o, and the effect on other variables is determined by the time series—results differ if X is held fixed throughout the resulting series, or merely 'shocked' at one time. Algorithms for computing the equilibrium effects using linear cyclic graphical models were developed in the engineering literature.

4. *Logistic regression models*. Suppose Y is a binary variable (say with values 0 and 1) and X a continuous variable. The *odds* that $Y = 1$ are $\Pr(Y = 1)/1 - \Pr(Y = 1)$ and the *log odds* or *logit* is the logarithm of that ratio. In logistic regression models, the logit of Y is set equal to a linear function of X plus noise, that is:

(2) *log* $\mathrm{pr}(Y = 1)/\mathrm{pr}\,(Y = 0) = aX + \varepsilon$

Versions of logistic regression are among the most widely used statistical models, especially in epidemiological contexts where causation is at issue; the model suggests that an intervention that changes X will change the *difference* (or the ratio) in the logs of the probabilities of the two values of Y.

5. *General Additive Models* allow for dependent variable Y to be any smooth additive function of other variables plus noise, e.g. $Y = a \ln(X^2) + b\, e^{cz} + \in$. Their causal interpretation is generally straightforward.

6. *Polynomial Regression Models* allow Y to depend on any polynomial function of other variables, plus noise. Again, the interpretation as causal models is straightforward.

7. *Time series models*. Consider an indexed set of vectors V_t with t ranging over the integers or the positive integers, and let there be a joint probability distribution that is stationary—the marginal distribution is the same for all times t. For such systems, and more specifically for linear systems in which the co-variances do not change with time, Granger (1969) proposed

that the series X_t is a cause (now called a *Granger cause*) of the series Y_t with respect to the remaining variables provided the expected value of Y_t conditional on $V_t \setminus \{Y_{t-}, X_t -\}$ is not equal to the expected value of Y_t conditional on $V_t \setminus Y_t$, where Y_t- and X_t are all variables for Y and for X, respectively with indices smaller than t. The idea is a generalization of linear regression and related to Suppes (1970) more general proposal for understanding causal relations as conditional probability relations subject to a time constraint.

Granger causation need not be causation if, for example, the difference in expected values is due to unrecorded common causes.

8. *Categorical variables* (Bishop, Fienberg, and Holland 1975) were the subject of various attempts over many years to provide a family of probability distributions and representations that could be estimated and could be naturally interpreted as specifying causal relations. One influential proposal is the log linear model, which specifies, not the dependence of variables on one another, but the joint probability of an assignment of values to categorical variables. With a system of variables of $n \times m$ values and a joint probability distribution one can associate an $n \times m$ table, with each cell of the table representing the probability that a case exhibits the combination of values in that cell. The log linear model treats the natural log of the cell probability as a linear function of an undetermined parameter raised to the power of each variable and each conjunction of variables. For example, for two binary variables, A, B, the natural log of the probability of a case lying in cell ij is $\ln (f_{ij}) = \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$. The parameter estimation problem is to determine the values of the λ variables. The log linear model has been given a causal interpretation for certain spatial statistics (Moore 2001) but has no evident general causal interpretation, although attempts have been made to give it one (Goodman 1978).

In its most general form, the causal Bayes net model for categorical variables assumes a multinomial distribution of the variables, and a directed acyclic graph of causal relations. The parameters of the distribution are simply the probabilities of each value of each variable conditional on each assignment of values to its parents—its direct causes—in the graph. Such models have the causal interpretations described in more detail below. More specialized parametric families of probability distributions for causal Bayes nets have also been used. For binary variables, a consistent parameterization (Pearl 1988; C. Glymour 2003) is obtained by treating each variable as a Boolean function of its parent variables, with each parent variable multiplied by a Boolean parameter, and taking probabilities of both sides.

3. PARAMETER ESTIMATION

While in principle every method of searching for causal explanations based on sample data might be structured as a form of parameter estimation, the most common view of causal inference in statistics is that it involves estimating unknown values of numerical parameters that measure the strengths of potential causal relations that are themselves specified a-statistically—from prior theory or from experimental design. In the simple 'model'

(3)   *crop yield/acre = α tons of fertilizer/acre + β tons of water applied/acre + χ + ε*

the aim is to estimate the unspecified values of α, β, and χ under various assumptions, for example that ′ is a normally distributed random variable and that values of α, β, and χ are the same for all cases ('fixed effects') or vary from case to case ('random effects'), etc. An estimator is simply a *statistic*: that is, a function from sample properties to some definite range of mathematical objects, in the present case to the real numbers that are the possible parameter values. In the simplest cases, estimates of the parameter values and assumptions made about the joint distributions of variables—in our example, ∈, tons of fertilizer/acre and tons of water applied/acre—determine a sampling distribution for a statistic—that is, for any given sample size, the probability distribution of values of the statistic among samples of that size. The oldest statistic of this kind is Legendre's least squares.

A large body of statistical results concerns which estimators meet various intuitively motivated criteria that can be assessed without knowledge of the true value of the parameter (Lehmann 1998). One criterion is 'consistency', which means, roughly, convergence of estimates to the true value of the quantity estimated. There are importantly different exact definitions. In all definitions below, α is a parameter or vector of parameters, *a* is its actual value, and $\bar{\omega}$ (α, N) denotes the value of the estimation function ω for α applied to a sample of size N and $\bar{\omega}$ (α, *d*) the estimate of α from *d*, the true probability density, and Pr is a probability based on *d*.

(1) Fisher Consistency: Given the (true) density *d* for the observed variables, when the value of parameter α is *a*, $\bar{\omega}$ (α, *d*) = *a*.

(2) Pointwise Consistency:

$$\forall \delta > 0, \forall \varepsilon > 0, \forall\, a, \exists\, n\, \forall\, N \quad > \quad n,\ [\mathrm{pr}\!\left(\bar{\omega}(\alpha, N) - \mathbf{a} > \delta\right) < \varepsilon|$$

(3) Uniform Consistency:

$$\forall \delta > 0, \quad \forall \varepsilon > 0, \ \exists\, n\, \forall\, N > n, \forall\, \mathbf{a}[\mathrm{pr}\!\left(\bar{\omega}(\alpha, N) - \mathbf{a}| > \delta\right) < \varepsilon].$$

All these consistency criteria have 'weak' versions that allow the estimator to be a partial function—that is, on some data the estimator can pass, and can continue to pass no matter how large the sample size, but must eventually provide an estimate as the sample size increases without bound. Analogous criteria apply as well to hypothesis testing and to procedures that search for graphical causal models, discussed below.

Uniform Consistency, but not Pointwise or Fisher Consistency, entails that confidence intervals for the estimates can be constructed that converge to zero width as the sample size increases without bound. For some statisticians, no estimation procedure is acceptable unless it satisfies Uniform Consistency, a requirement that sometimes excludes all possible estimators.

None of these consistency criteria suffice to distinguish among many possible estimators, and other desiderata are therefore imposed where they can be. For example, the mean squared error of estimates can be divided into a term representing the expected absolute value of the difference between the true value and the estimated values—the square of the bias—and a term representing the variance of the estimates (the variance of the estimates, that is, that

would be obtained on samples of the given size obtained from the true distribution.) A common requirement is that an estimator be unbiased and among unbiased estimators, have the minimum variance. A popular alternative, Fisher (1990), is maximum likelihood: that the estimated value be that for which, among the alternative possible values, the observed sample is the most probable. An enormous literature studies the applicability and interconnections of these and related criteria.

An increasingly popular alternative is Bayesian estimation of parameters, which, starting with a probability distribution over the values of the parameters, computes a probability distribution conditional on the observed sample (Lee 2004). Bayesian estimation, long merely a toy because of the difficulty of actually computing posterior distributions, has been made practical by simulation methods that allow such estimates for small samples (Casella and George 1992) and by an easily computed asymptotic formula, the Bayes Information Criterion (Schwarz 1978) that in many cases provides good approximations to the posterior probability for large samples. Disputes over the various consistency criteria above are of little relevance to Bayesians, who have weaker requirements, for example, that the set of values of the parameter for which the posterior probability converges, in the pointwise sense above, has probability 1, or that the expected error converges to 0.

Parameter estimation has an underdetermination problem. *Identifiability* fails when more than one assignment of parameter values determines the same marginal probability distribution over observed variables. Identifiability typically fails for parameters relating variables that are *confounded*, that is, jointly influenced by one or more common unobserved variables. When X is thought to be a confounded cause of Y, a standard solution is to find a *instrumental variable Z* that is thought to influence Y if at all, only through X; in some distributions this permits consistent estimation of the influence of X on Y, for example, of α in equation (3). The instrumental variables technique works only for special forms of dependency and probability distributions; while it holds for systems of binary variables parameterized as 'noisy or gates' (e.g. $\Pr(Y) = \Pr(a\,X \oplus b\,Z)$, where Y, X, Z, a, and b are Boolean and $\oplus$ is Boolean addition) (Glymour 2003), it fails, for example, for categorical variables distributed according to a multinomial distribution, although bounds on probabilities may be estimated (Galles and Pearl 1995).

Statistical literature and practice contain various other ad hoc or heuristic rules for avoiding confounding, in particular advice to condition on any variable found to be associated with both of two variables thought to be causally related, and to stop conditioning on new variables when the association under study does not change much. While widely used, this recommendation is not generally sound and can result in increased error compared to estimates with a smaller or larger set of co-variates.

Finally, some distributions and their parameters are regarded as 'not causal' for good reason. For example, in a linear system with normally distributed variables, the variables may be transformed, or 'standardized', by setting, for all variables X, $X_s = (X - {}_{\mu X})/{}_{\sigma X}$, where ${}_{\mu X}$ is the mean of X in the sample and ${}_{\sigma X}$ is the sample standard deviation. The result is that linear coefficients become correlation coefficients, but do not predict the effect of an intervention that produces a unit change in a causal variable in any other sample governed by the same causal process but with different noise variances.

4. HYPOTHESIS TESTING

Parameter estimation is often an implicit step within testing a causal hypothesis expressed by an equation, such as (1), assuming a family of probability distributions characterized by values of the parameters in the equation. For example, from a maximum likelihood estimate of parameter values a sampling distribution of some statistic is obtained and the probability that the value of the statistic lies in some tail or tails of the distribution is computed, ideally in conjunction with the probability of the same tail membership of the statistic as a function of alternative values of the parameters—essentially, the *power function* of the test. Depending on the school of statistics, the use of the test is to reject the hypothesis, or not (Neyman–Pearson), or to regard the hypothesis as confirmed, or not (Fisher). There are as many tests of a model as there are relevant statistics, and a statistical standard is to search for 'uniformly most powerful' tests and recommend their use when they exist.

Philosophical commentators (Mayo 1996) have emphasized that reliable inquiry requires some further criterion of severity of a test, or body of tests, in excluding alternatives, where the severity of a test of a hypothesis is, roughly, a function of the probability that the test would result in a worse fit to the sample data were the hypothesis false; Mayo proposes that experiment with hypothesis testing be extended to all of the assumptions of a 'model' such as (3), and to the sampling procedure, with the hope that unique explanatory features will eventually be identified. Assumptions about the distribution family may be tested, as may assumptions that the probability distribution for values of variables are the same for all sample units, and independent for each sample unit (i.i.d. sampling), and so on. A careful formulation of severity criteria and a more detailed discussion of foundational issues in hypothesis testing may be found in Mayo (1996). This perspective is illustrated vividly in work by Spanos (2007), who develops a heuristic search procedure based on a series of hypothesis tests for finding the appropriate family of probability distributions, and for detecting and correcting for statistical dependencies between units in a sample (usually called, rather misleadingly, autocorrelation).

5. ESTIMATING INTERVENTIONS AND GRAPHICAL MODELS

The graphical causal model framework exploits directed graph representations, representations that have a long history, reviewed in Pearl (2000). Variables are represented as nodes in a directed graph, and a directed edge, $X \to Y$, is the claim that X is a direct (relative to other variables represented in the graph) cause of Y. The diagrams are useful as a psychological aid, but so far as estimation is concerned, the key tools are consequences of 'factorizations' of the joint probability density on values of a set of variables. In graphs without cycles—paths of directed edges with the same orientation that begin and end with the same variable—the factorization is implied by the Markov assumption: each variable X, in a directed graph is independent in probability of variables in the graph that are not direct or indirect effects of X, conditional on the direct causes of X in the graph. Formulations of the Markov assumption (Kiiveri and Speed 1982) for causal systems emerged from studies in the late 1970s of factorizations of distributions. Let V be a set of variables V_i, the vertex set for a

directed acyclic graph (DAG) G, and for each variable Vi in V, let Par(V_i) be the set of variables with edges in G directed into V_i. Let Pr be a probability distribution or density satisfying the Markov assumption for G. Then for all sets v consisting of one value, v_i, for each variable V_i in V,

$$(4) \quad \mathrm{Pr}(V = v)) = \Pi_i \mathrm{Pr}\left(Vi = v_i | \mathrm{par}(V_i)\right)$$

The Markov factorization is a *necessary* consequence of either of the following:

   a. the value of each variable is determined by the values of its parents and zero indegree variables are jointly independent;
   b. the joint probability distribution is the marginal of a probability distribution satisfying the Markov assumption for a directed graph without cycles, zero indegree variables are jointly independent; and some (possibly empty) set of variables, each of which has a single direct effect in the graph and is the effect of no variable in the graph, is marginalized out.

The causal graphical framework aims (1) to enable the estimation of the probability of any represented variable conditional on values of any other variables represented; (2) to enable the estimation of the probability of any represented variable conditional on any hypothetical intervention that forces new probability distributions on other variables. Our concern here is with the second, causal, aim. The probabilities of variables in a graphical model can be interpreted either as actual or hypothetical population frequencies, or as propensities or chances attached to individuals, propensities that may differ from the value a variable actually has for the individual.

Pearl, Geiger, and Verma (Pearl 1988) and Lauritzen and his collaborators (Lauritzen 1996), provided algorithms that decide whether the Markov assumption applied to a directed acyclic graph implies any particular conditional independence relation. Pearl's version has been more popular: V_1 and V_2 are *d-connected* conditional on set Z if and only if there is a sequence of edges between V_1 and V_2 such that every vertex touched by the sequence and having two of the sequence edges directed into it (every *collider* on the sequence) is in Z or is the source of a directed path leading to a member of Z, and no other vertex touched by the sequence is in Z. Vertices are *d-separated* with respect to Z if they have no d-connecting path with respect to Z. The Markov assumption applied to a directed acyclic graph implies that in *every* distribution satisfying the assumption for the graph, two vertices are independent conditional on a set of other variables if the vertices are d-separated conditional on the set. Many of the notions that are otherwise explained in terms of correlations of various kinds, or their absence, are explicable in terms of d-connection. For example Z is an *instrument* for the effect of X on Y provided (i) Z, X, and Y are not pairwise independent; (ii) there is no edge from X to Z; and (iii) every edge sequence that unconditionally d-connects Z and Y touches X. The d-separation relation also characterizes independence and conditional independence relations in systems in

which the noise terms for some variables are specified to be correlated, without causal explanation, and it also necessarily characterizes those relations implied (for all non-zero values of linear coefficients) by linear systems represented as *cyclic* graphs under the two circumstances listed above as sufficient for the Markov condition.

Given an acyclic causal graph $G = <V, E>$, V a set of random variables and E a set of directed edges, and a probability distribution Pr, Markov for the graph, an intervention on V in V can be represented by an extension of G with a new variable I_V with at least two values and an edge directed into V and no edges into or out of I_V, and a probability distribution Pr*, Markov for the extended graph, such that Pr* conditional on one value i of I_V is equal to Pr, and conditional on at least one other value j of I_V is equal to Pr with a new factor Pr*(V) substituted for the original Pr (V) in the factorization. In general, if the factorization of V depends on Par(V), in the new distribution the factorization of V depends on Par (V) $\cap$ I_V The representation allows the computation, given the probability of V conditional on $I_V = j$, of the probability of any other variables produced by an intervention (Spirtes et al. 2001). For the special case of interventions that 'break' edges into V—i.e. that in the factorization of the original graph remove the dependence of V on other variables, Pearl (2000) provides an algorithm for the computation when the causal structure is known but may contain latent variables, extending an algorithm of Spirtes et al. (2001). Either procedure can be applied when causal and probabilistic input is incomplete, and Spirtes's form can do so even when the intervention alters, but does not remove, the conditional probability of V on Par(V). Spirtes's algorithm can give 'not computable' as an outcome, and it is known that the procedure is not maximally informative, but Pearl's has been shown to be. Woodward (2003) has argued that the edge-breaking sense of intervention is fundamental to the notion of causation. The Markov condition and associated algorithms yield results about estimation of intervention effects that hold for *every* probability distribution Markov for *any* DAG. Further prediction results may hold for particular families of probability distributions; the consequences of restrictive distribution assumptions are not very well explored.

6. THE COUNTERFACTUAL FRAMEWORK

A more influential framework in contemporary statistics (Rubin 1977; Robins 1986) analyses causal dependency as a counterfactual relation, roughly in the sense of Lewis (1973), although philosophical logicians seem not to have been read by the statistical community. The goal of inference is taken to be the effect of treatment assignment T = t on outcome O for unit u, defined to be the difference between the actual value of the outcome for u and the outcome u would have had under an alternative treatment assignment, T = t'. 'Treatment assignment' is sometimes a misnomer, since the counterfactual framework is meant to be applied to non-experimental data as well as to experimental, and merely means whichever variable of a pair is considered to be the potential cause. ('Treatment assignment' is distinguished from 'Treatment' in the epidemiological literature because patients do not always do as told.)

Since the alternative treatment is not given, the outcome that would have obtained had an individual been given an alternative treatment is not observed. The distinguishing idea of the

counterfactual framework is to introduce for each outcome variable and each alternative treatment assignment a 'counterfactual variable' whose value is, for each individual in the sample, the value the outcome would have had for that individual if the alternative 'treatment' had been given. Contrasts between the outcome on one treatment assignment and on another treatment assignment then become a problem in estimating models with unobserved, counterfactual variables. This kind of estimation problem has no unique answer for the individual case. Estimation of average influences requires assumptions about the uniformity of dependencies across individuals (or the random distribution of dependencies independently of the variable values) and the absence of confounding variables influencing the putative cause and the putative effect. Statisticians using the framework tend to report 'propensity scores' showing how the contrast depends on the values of parameters in the model representing counterfactual dependencies and confounding influences.

Models developed within the counterfactual framework implicitly use the Markov assumption, which suggests these models have graphical model equivalents that eschew counterfactual variables, as various authors have (e.g. Pearl 2000) have argued. There are differences. The counterfactual framework does not allow counterfactual variables that range over values the *treatment assignment* (or other upstream variable) would have had were the effects to have been different. Thus, unlike the graphical causal model framework, in order to specify the relevant model variables the counterfactual framework requires prior assumptions as to which variables are potential causes of which others. For this and other reasons, work in the counterfactual framework avoids automated search procedures.

7. SEARCH FOR CAUSAL EXPLANATIONS

Experimental settings in which one or more variables are manipulated by the experimenter provide restrictions on plausible causal hypotheses. If the value of X in each case is determined by the experimenter, then values of X and of potential effects, Y, of X are not confounded by common causes, and Y is not an effect of X. The inquiry is reduced to estimating whether X has any effect on Y, and, if so, some measure of the strength of that effect. Early in the twentieth century, Fisher (1990) made popular designs that randomize the values assigned to variables under experimental control, both avoiding confounding and allowing statistical tests of the hypothesis of no effect, and estimates (as by ANOVA or regression) of the strengths of influences. Fisher also introduced strategies for making statistical inference more efficient—both in an informal and in a technical sense—by various sampling and control methods. For example, in estimating which of two kinds of shoes wear longest on boys, simple randomization would assign a pair of shoes of one or another kind at random to a representative sample of boys. But, on average, boys might wear out shoes on their right feet more quickly than on their left, and boys vary in how rapidly they wear out shoes. Estimates of the difference would be improved by randomly assigning, for each boy, one shoe of one type to one of the feet of a boy, and the other type to the other foot of the same boy and estimating the average across all boys of the individual differences in wear. Recent work in the graphical causal model tradition has shown that when the aim is to determine all causal relations among a set of variables, strategies that simultaneously and independently randomize multiple variables reduce exponentially the number of experiments required.

Related results also suggest improved estimation of causal effects through such strategies (Eberhardt, Glymour, and Scheines, 2005).

Some Bayesian statisticians dispute the necessity of randomization, and the appropriateness of relying on randomization to remove confounding. A random sample may, by chance, be very unrepresentative of the population from which it is sampled.

Even with experimental manipulation, causal inference can be difficult. The treatment and its effects may, for example, cause some units in an experiment to drop out or not comply with the experimental design (e.g. cells die, mice die, people stop taking their drugs, people drop out of a long-term experiment), resulting in an unrepresentative ('biased') sample. Experiments may seek simultaneously to estimate the effects of a variable on multiple outcomes, but the several outcome variables may influence one another, or there may be unmeasured confounding variables that influence the outcomes and are not removed by the randomization—because the *outcome* variables are not randomized. In these respects, causal inference from experiments shares problems with causal inference without experimental controls.

Without experimental controls, search for causal relations from samples may be viewed essentially as a kind of estimation problem in which hypothesis testing may (or, as in Bayesian procedures, may not) be a tool or step. Causal estimation may be done in steps, first estimating the graphical causal structure and possibly the functional form of dependencies, then estimating parameters, or the functional form may be separately estimated (or assumed) and the estimation of parameters and graphical structure estimated simultaneously. The estimated causal structures are themselves hypotheses that can in most cases be subjected to statistical tests. The mathematical questions are of the same kind as in conventional statistical estimation: under what assumptions do which kinds of search procedures (i.e. estimators) have which kind of desirable statistical and computational properties connected with (probably) finding the truth?

Despite the parity of reasoning between estimation and search, a long and influential tradition in statistics has deprecated model search as 'fishing expeditions' or 'ransacking'. One intelligible thought behind the slogans was that using data to develop and test a statistical model would lead to 'overfitting', meaning that estimates of parameter values obtained from the data also used to obtain the model would in general not agree with estimates of the same parameters obtained from new data drawn from the same probability distribution (because the model obtained would sometimes be wrong, typically containing too many parameters). Statistical writers distinguished 'confirmatory' statistical analyses, usually meaning those that issued in a well-defined test of the hypothesis on data not 'used' in formulating the hypothesis, and those analyses resulting from data-driven search that had no such test. Bayesian statisticians have been especially concerned about 'double counting', that is, using the same datum in forming a prior probability distribution and in calculating a posterior distribution. These objections are now often addressed in practice by holding out a sample of data for testing, or by repeating model search and parameter estimation with subsets of the original data and testing the model and estimates on the remainder of the original data (usually called *cross-validation*). Further, it was for a long time quite unclear what mathematical objects could represent causal relations, and without such objects model search

could not be treated mathematically as estimation. The graphical representation of causal relations and formalization of the Markov condition have largely, but not entirely, resolved that problem. One could reasonably doubt that automated procedures could be devised that would substitute for knowledge of a domain and human consideration of the data. For example, the fact that test scores of students on an examination can be put in a series in which there are improbable sequences of similar scores would suggest nothing to a computer, but to a human, knowing that the series order is the seating order, it is evidence of cheating. The doubt is well-founded, but that does not mean that systematic, automated search procedures cannot help in discovery.

From early in the twentieth century, statisticians recognized that in many problems the number of potential alternative causal explanations for a body of data is infinite, or at least too large to survey explicitly, and that the fact that a particular model is not rejected by a test provides no guarantee that a particular alternative model is the true one, or at least relevantly closer to the truth. Further, the implicit statistical criterion of success involved succeeding—converging to the truth—regardless of what the truth might be, without sometimes having to say 'don't know'. This meant that causal parameters could never be pointwise estimated unless the graphical structure and parametric family were already known, and it was far from clear how such knowledge could be acquired systematically and reliably. There are a great many hazards to correct causal inference from non-experimental data, for example: (1) missing values of variables for cases; (2) unmeasured confounding variables; (3) measurement errors; (4) sample selection bias; (5) autocorrelation, in which values of variables for a sample unit influence values of variables in other sample units; (6) probability distributions and functional dependencies that are not among the familiar examples; (7) samples that are formed of sub-populations with distinct probability distributions and even distinct qualitative causal relations; (8) the data may be described best by a *cyclic* graph, and for reasons noted above, the causal content of such models is ambiguous and their discovery from data alone seemed implausible; (9) sometimes the causal relations of interest are among variables that are not measured, but whose effects or manifestations are measured, and it seemed implausible—some claimed impossible (Bartholomew and Knott 1999)—that data-driven methods could provide the information required.

The development of the formalism of graphical causal models prompted a burst of research beginning around 1990 on computerized search methods for which proofs of convergence to correct information could be provided. The Markov assumption alone is insufficient for the existence of any consistent estimators of causal relations, and further assumptions are needed to form a subspace of possible models for which search is possible. One widely used assumption is *Faithfulness*: all conditional independence relations in the distribution are implied by the Markov assumption applied to a DAG. Faithfulness was later shown to hold probability 1 for DAGs with smooth measures on the parameters of linear models or the parameters of categorical variable models (Spirtes et al. 2001). Markov, Faithfulness, and samples that are independently and identically distributed have been proved to be sufficient for pointwise consistent inference to features of causal graphs in a variety of circumstances: (a) for acyclic causal structures, when there are no unrecorded confounding variables or 'correlated errors'; (b) for linear, cyclic causal structures when there are no unrecorded confounding variables or correlated errors; (c) for acyclic causal structures when there are

unrecorded, and unknown (before data analysis), confounding variables or correlated errors; (d) for identifying sets of measured variables that share a single unmeasured common cause; and (e) for estimating features of the causal relations among the latent variables identified as in (d). The same algorithm that suffices for (c) is also pointwise consistent when there is sample selection bias. These procedures do not typically identify a unique directed graph of causal relations, but rather features (e.g. a directed edge, or a directed path, etc.) common to all members of a set of alternative graphs that might explain the data. More recent work has shown that unique DAGs for linearly related, non-normally distributed systems without latent variables can be consistently identified from i.i.d. data provided measured variables are not deterministic functions of one another. (Shimizu et al. 2006). Faithfulness is not required—effects due to different causal pathways can perfectly cancel and the causal structure can nonetheless be fully recovered. These methods have been combined with procedures for identifying latent variables to estimate a unique DAG among unrecorded common causes.

Linear autoregressive time series can be given the form $y_n = x_n + \sum_j a_j y_{n-j}$ where j ranges over some specified number of previous time steps, the a_j are real constants, and x_n is a 'random shock' at time step n. In the multivariate version each variable may depend on a linear combination of time delays of other variables. Moving average time series make y_n a linear function of past random shocks plus a current shock. When both sorts of dependencies are present, the system is called an autoregressive moving average, or ARMA model. Time series models have an autocorrelation between any two variables for any specified time difference, or lag—the correlation of y_n with x_{n-j} for some fixed j. Partial autocorrelation for a given time difference is autocorrelation conditional on values of the variables between the lags.

A standard search procedure for time series is owed to Box and Jenkins (1970). An empirical series is tested, and if necessary adjusted, for stationarity—the joint distribution of the variables at a time step must be independent of the time step, or, for Gaussian processes, the co-variance of variables must be independent of time. Patterns of autocorrelation and partial autocorrelation are then used to determine which variables influence which others at which lags, and the values of parameters are estimated by standard statistical procedures. It can happen that two series are not stationary, but some linear combination of them is. If, of two such series, the two derivative series obtained by taking the difference of values between each time step are both stationary, then the two original series are said to be co-integrated. One analogy for co-integrated series is a man walking a dog on a leash. As the man pulls the dog and the dog pulls the man, each of their trajectories is irregular, but the average of their positions follows a much smoother trajectory.

A simplified procedure is sometimes used to find Granger causes: each pair of variables is regressed in each direction, controlling for other variables at a large number of lags; significant regressions are taken to indicate a causal relation. Many complexities arise in searching for multivariate causal time series: dependencies may be non-linear, distributions non-normal, stationarity may not hold, a series of unobserved common causes may exist, and, because many time series are in discrete steps, 'contemporaneous' causal processes may occur between the time steps. The last problem has been essentially solved for contemporaneous causes by first regressing each variable on all previous time steps of all variables, and then

applying graphical search algorithms referenced above to the residuals (Demiralp and Hoover 2003; Moneta and Spirtes 2006). The procedure allows unobserved contemporaneous common causes. Time series problems remain very much open.

8. CAUSAL INTERPRETATION AND CAUSAL PUZZLES

The statistical literature has developed various standard puzzles about probability and causality that sometimes take on a didactic role.

1. *Mistaken mechanisms*. The overall rate of acceptance of female applicants to graduate programmes at UC Berkeley was lower than the overall rate of acceptance of male applicants, prima facie evidence of discrimination against women. But not really: within each department, women were accepted at the same rate as men, but women tended more often to apply to programmes for which admission was more competitive.

2. *Zero correlation*. The correlation between per capita foreign aid that nations receive and the proportion of a nation's population living on less than one dollar a day, is zero. It does not follow and is not true that increasing foreign aid to any particular nation would not decrease poverty in that nation. What does follow, or at least is suggested, is that a marginal increase in the amount of foreign aid, if distributed among nations according to current mechanisms, would not decrease extreme poverty.

3. *Correlation and aggregation*. Suppose the RNA is extracted from each cluster of cells and the concentrations measured, for many cell clusters. Suppose the concentrations of two molecular species of RNA are strongly correlated, and remain correlated conditional on a third measured species correlated with the first two. Each RNA species can be mapped to a 'reading frame'—a gene fragment from which that RNA molecule is produced by transcription. Does the correlation mean that transcription of one of the genes influences transcription of the other, or that there is an unrecorded common cause? It does not, because the measurements are effectively of *sums* of concentrations over many cells. If the relations between transcription of one gene and transcripts of other genes is non-linear, as it is thought to be in some cases, then conditional independence relations among concentrations at the cellular level can become conditional *dependence* relations among aggregated concentrations.

4. *The Monty Hall Problem*. Monty Hall places a pile of money behind one of three doors. A contestant arrives and chooses a door. If the door the contestant chooses actually hides the money, Monty opens one of the other doors at random; otherwise Monty opens the door that the contestant did not choose and that does not hide the money. Seeing the open, empty door, the contestant now has the option of changing his selection or staying with his original choice. He wins the money if the door he finally chooses, whichever it is, hides the money. Should he switch doors or stand pat?

In repeated plays, contestants will win twice as often if they switch doors. The argument is simple: there was a 2/3 chance the original choice was incorrect, and subsequently removing an alternative does not change that fact. A similar result would obtain with 100 doors: one would win 99 times out of a hundred by switching after 98 randomly chosen doors were opened. The connection with causality is as follows: Monty's choice of where to put the money and the contestant's original choice of doors are independent variables, each with three

possible values. Each of these variables influences which door Monty opens, another variable with three values. Independent variables that mutually influence a third variable are (almost always, assuming faithfulness) dependent conditional on the value of the variable they both affect, and that holds in this case. Given the information about which door Monty opens, the contestant's original choice of doors provides information about where Monty put the money.

5. *Simpson's Paradox*. Simpson (1951) produced an imaginary case in which the story suggests that neither X nor Y influence Z, X and Y are independent, but X and Y are dependent conditional on Z. His imaginary example produced a considerable statistical (and, eventually, philosophical) literature on reversals of association by conditioning on other variables. From the point of view of graphical models, Simpson's example is a contrived story with an unfaithful distribution, but the changes in association, including reversals of sign, resulting from conditioning is a fundamental problem that besets causal inference from regression.

6. *Lindley and Novick's Puzzle*. Suppose we have the data for variable Y, with values y1 and y2, and variable X with values x1 and x2, for two groups, one with value z1 for variable Z and the other with value z2 for Z (Fig. 23. 1). Taking the sample as representative of a joint probability distribution, none of the variables are independent of any others. There is, however, something odd about the conditional probabilities: $p(y1|x2, z1) > p(y1|x1, z1)$, and $p(y1|x2, z2) > p(y1|x1, z2)$, but $p(y1|x2)$

Z, then no matter which value of Z we condition on, also conditioning on x2 gives y1 a higher probability than does also conditioning on x1. But if we do *not* condition on any value of Z, conditioning on x1 gives y1 a higher probability than does conditioning on x2. No pair of the variables is independent conditional on the third.

Lindley and Novick (1981) pose two different ways these data could have been generated:

1. A medical experiment: X = treatment; x1 = treated; x2 = not treated; Y = outcome; y1 = recovered; y2 = did not recover; Z = sex; z1 = male; z2 = female.
2. An agricultural experiment: X = variety of plant; x1 = white, x2 = black; Y = yield; y1 = high, y2 = low; Z = height of plant; z1 = tall; z2 = short. They raise these questions: if you want to produce the best medical effect, to whom if anyone should the treatment be given? If you want to produce the best yield,

z1	y1	y2	Total
x1	18	12	30
x2	7	3	10
Total	25	15	40

z2	y1	y2	Total
x1	2	8	10
x2	9	21	30
Total	11	29	40

Fig. 23.1

what variety of plant should be planted? They answer: for the medical case, no treatment should be given, that is, x2 is the 'non-treatment' of choice, because it has better recovery probabilities for males and better recovery probabilities for females. In the agricultural experiment, white plants should be grown (i.e. x1) because, overall, it has a better probability of high yield.

Their explanation is that under the second interpretation, but not the first, the cases in the data are 'exchangeable'. The idea of an exchangeable probability distribution is simply that for any finite ordered sample, the probability of obtaining that sample, given the sample size, is the same as the probability of obtaining a sample of the same size with any permutation of the given ordering. The exchangeability explanation is mysterious.

The two different stories, one medical and the other agricultural, naturally lead to different causal interpretations of the data, and the different causal interpretations suggest different effects of interventions (Meek and Glymour 1994; Pearl 2000). In the medical case the task is to estimate the relative effects of treatment versus no treatment in the experiment, and then use that information to recommend how the general population should be treated, or not treated. Assume that in the experiment someone's sex is not caused by the medical treatment. Sex and treatment, *Z* and *X*, are not independent in the data—more men received the treatment than did women. The dependency must then either be due to chance in selection of subjects, or due to the influence of sex on which patients were treated and which were untreated, or due to the influence of something unknown on both sex and treatment. The causal structure of the experiment in the two alternative later cases is one of those shown in Figs. 23.2 and 23.3.

The case of Fig. 23.3 is implausible if we think of sex as the *biological condition* of a person, but not if we think of sex as the biological condition of a *subject selected for the experiment*. In either case we know from the Markov Assumption how to compute the probability of recovery, y1, from an intervention—a forced value of x1 or of x2—in a system with this description. We eliminate the association between treatment choice and outcome due to the common cause (sex in Fig. 23.2; unknown in Fig. 23.3), and use the remaining association as our estimate of the effect of treatment choices on recovery. We can do that by conditioning on Z, on the sex of the subjects. We compute the probability of y1 conditional on x1, for example, and conditional on Z. The probability of y1 conditional on x1, or respectively on x2, is of course different for different values of Z, for males or females, but x2 is better in both cases.
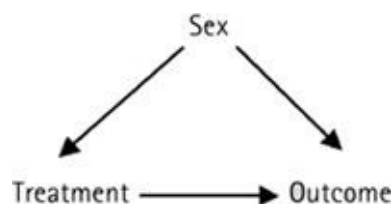
Fig. 23.2
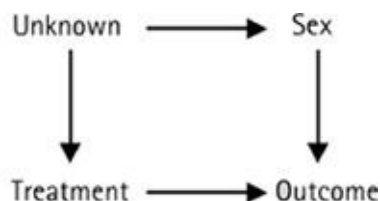
Fig. 23.3
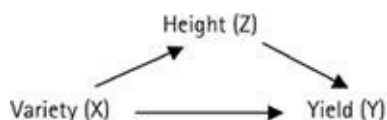
Fig. 23.4
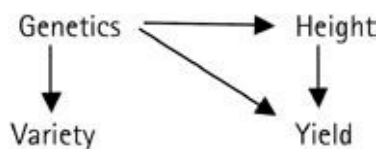
Fig. 23.5

In the agricultural interpretation of the experiment, heights of plants are correlated with variety of plant. That is most plausibly because the plant's variety influences its height, or something else—genetics, say—influences both the variety and the height of the plant, but the height of the plant doesn't influence the variety. (Both mechanisms are possible, of course.)

In recommending a planting policy, we know the variety of seed to be planted, black or white, but we do not know when we plant whether the plant will be short or tall. If in fact the plant variety influences height, as in Fig. 23.4, then the variety influences the yield through two mechanisms, one direct, and the other through height. In that case, to assess the influence of variety on yield, we *should not* condition on height. To do so would be to discount one of the paths by which variety influences yield. So we find Lindley and Novick's conclusion. Various complications arise if we think of genetics as a common cause of variety and height, as in Fig. 23.5.

A variety of other difficulties in the causal analysis of epidemiological data are clearly explained from the perspective of graphical causal models in M. M. Glymour (2006).

FURTHER READING

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer.

Bishop, Y., Fienberg, S., and Holland, P. (1975). *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT.

Fienberg, S. (1975). *The Analysis of Cross-Classified Categorical Data*. Cambridge, Mass.: MIT.

Fisher R. A. (1990). *Annual Proceedings of the Conference on Uncertainty in Artificial Intelligence. The Journal of Machine Learning Research*.

Glymour, C. and Cooper G. (eds.), *Computation, Causation and Discovery*, Cambridge, Mass.: MIT.

Spirtes, P., Glymour, C., and Scheines, R. (2001). *Causation, Prediction, and Search*. Cambridge, Mass.: MIT.

Pearl J. (2002). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge: Cambridge University Press.

REFERENCES

BARTHOLOMEW, D., and KNOTT, M. (1999). *Latent Variable Models and Factor Analysis*. 2nd edn. London: Edward Arnold

BISHOP, Y., FIENBERG, S., and HOLLAND, P. (1975). *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT.

BOX, G. E. P., and JENKINS, G. M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.

CASELLA, G., and GEORGE, EDWARD I. (1992). 'Explaining the Gibbs Sampler'. *The American Statistician* 46: 167–74.

DEMIRALP, S., and HOOVER, K. (2003). 'Searching for the Causal Structure of a Vector Autoregression', *Oxford Bulletin of Economics* 65: 745–67.

EBERHARDT, F., GLYMOUR, C., and SCHEINES, R. (2005). 'On the Number of Experiments Sufficient and in the Worst Case Necessary to Identify All Causal Relations Among N Variables', in F. Bacchus and T. Jaakkola (eds.), *Proceedings of the* 21*st Conference on Uncertainty in Artificial Intelligence (UAI)*. Arlington, Va.: AUAI Press, 178–84.

FISHER, R. A. (1990). *Statistical Methods, Experimental Design, and Scientific Inference*. New York: Oxford.

GALLES, D., and PEARL, J. (1995) 'Testing Identifiability of Causal Effects', in P. Besnard and S. Hands (eds.), *Uncertainty in Artificial Intelligence*. San Francisco: Morgan Kaufmann, xi. 185–95.

GLYMOUR, C. (2003). *The Mind's Arrows*. Cambridge, Mass.: MIT.

GLYMOUR, M. M. (2006) 'Using Causal Diagrams to Understand Common Problems in Social Epidemiology', in J. M. Oakes and J. S. Kaufman (eds.), *Methods in Social Epidemiology: Research Design and Methods*. San Francisco: Jossey-Bass.

GOODMAN, L. (1978) *Analyzing Qualitative/Categorical Data: Log-Linear Models and Latent Structure Analysis*. Cambridge, Mass.: Abt Books.

GRANGER, C. (1969). 'Investigating Causal Relations by Econometric Models and Cross-spectral Methods', *Econometrica* 37: 424–38.

KIIVERI, H., and SPEED, T. (1982). 'Structural Analysis of Multivariate Data: A Review', in S. Leinhardt (ed.), *Sociological Methodology*. San Francisco: Jossey-Bass.

LAURITZEN, S. (1996). *Graphical Models*. Oxford: Oxford University Press.

—— (2001). 'Causal Inference from Graphical Models', in O. Barnsdorff-Nielsen, D. Cox, and KLUPPENLBERG, C. (eds.), *Complex Stochastic Systems*. London: Chapman & Hall, 3–107.

LEE, P. (2004). *Bayesian Statistics: An Introduction*. New York: Wiley.

LEHMANN, E. (1998). *Theory of Point Estimation*. New York: Springer.

LEWIS, DAVID (1973). *Counterfactuals*. Oxford: Blackwell.

LINDLEY, D., and NOVICK, M. (1981). 'The Role of Exchangeability in Inference', *Annals of Statistics* 9: 45–58.

MAYO, D. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

MEEK, C. (1995). 'Causal Inference and Causal Explanation with Background Knowledge', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. Philippe Besnard and Steve Hanks. San Mateo, Calif.: Morgan Kaufmann, 403–10.

—— and GLYMOUR, C. (1994). 'Conditioning and Intervening', *British Journal for the Philosophy of Science* 45: 1001–21.

MONETA, A., and SPIRTES, P. (2006). 'Graphical Models for Identification of Causal Structures in Multivariate Time Series', in *Joint Conference on Information Sciences Proceedings*. Paris: Atlantis Press.

MOORE, M. (2001). *Spatial Statistics*. New York: Springer.

NEAL, R. (2000). 'On Deducing Conditional Independence from *d*-Separation in Causal Graphs with Feedback', *Journal of Artificial Intelligence Research* 12: 87–91.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, Calif.: Morgan Kaufmann.

—— (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

—— and DECHTER, R. (1996). 'Identifying Independencies in Causal Graphs with Feedback', *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, ed. Eric Horvitz and Finn Verner Jensen. San Mateo, Calif.: Morgan Kaufmann, 420–6.

—— and ROBINS, J. (1995). 'Probabilistic Evaluation of Sequential Plans from Causal Models with Hidden Variables', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. Philippe Besnard and Steve Hanks. San Mateo, Calif.: Morgan Kaufmann, 444–53.

RAMSEY, J. ZHANG, J., and SPIRTES, P. (2006). 'Adjacency Faithfulness and Conservative Causal Inference', *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*. Arlington, Va.: AUAI Press.

RICHARDSON, T. (1996). 'A Polynomial-Time Algorithm for Deciding Equivalence of Directed Cyclic Graphical Models', in *Proceedings of the Twelfth Conference on*

Uncertainty in Artificial Intelligence', ed. Eric Horvitz and Finn Verner Jensen. San Mateo, Calif.: Morgan Kaufmann, 462–9.

ROBINS, J. (1986). 'A New Approach to Causal Inference in Mortality Studies with Sustained Exposure Periods—Application to Control of the Healthy Worker Survivor Effect', *Mathematical Modeling* 7: 1393–512.

—— SCHEINES, R., SPIRTES, P., and WASSERMAN, L. (2003). 'Uniform Consistency in Causal Inference', *Biometrika* 90: 491–515.

RUBIN, D. (1977). 'Assignment to Treatment Group on the Basis of a Covariate', *Journal of Educational Statistics* 2: 1–26.

SCHWARZ, G. (1978). 'Estimating the Dimension of a Model', *Annals of Statistics* 6: 461–4.

SMIMIZU, S. HOYER, P. MYARINEN, and KERMINEN, A (2006). 'A Linear Non-Gaussian Acyclic Model for Causal Discovery', *Journal of Machine Learning Research* 7: 2003–30.

SIMPSON, E. (1951). 'The Interpretation of Interaction in Contingency Tables', *Journal of the Royal Statistical Society, Series B* 13: 248–51.

SPANOS, A. (2007). 'Curve Fitting, the Reliability of Inductive Inference and the Error-Statistical Approach', *Philosophy of Science* 74: 1046–66.

SPIRTES, P. (1995). 'Directed Cyclic Graphical Representation of Feedback Models', *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ed. Philippe Besnard and Steve Hanks. San Mateo, Calif.: Morgan Kaufmann, 491–8.

—— and RICHARDSON, T. (1997). 'A Polynomial Time Algorithm for Determining DAG Equivalence in the Presence of Latent Variables and Selection Bias', *Proceedings of the* 6*th International Workshop on Artificial Intelligence and Statistics*.

—— GLYMOUR, C., and SCHEINES, R. (2001). *Causation, Prediction, and Search*. Cambridge, Mass.: MIT.

SUPPES, P. (1970). *A Probabilistic Theory of Causality*. Acta Philosophica Fennica 24. Amsterdam: North-Holland.

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

PART VI
CAUSATION IN PHILOSOPHICAL THEORIES

MENTAL CAUSATION

CEI MASLEN TERRY HORGAN HELEN DALY

To say that there is mental causation is to say that mind makes a difference. Beliefs, desires, feelings, and thoughts cause further beliefs, desires, feelings, and thoughts; and they also cause physical events, such as actions. That is, in addition to physical-to-mental causation there also seems to be mental-to-mental and mental-to-physical causation. Yet it is difficult to understand how there can be. In this chapter, we discuss the most popular current arguments against the existence of mental causation, and the troubles with some proposed solutions. We also promote one approach to causation that responds nicely to them all.

Mental causation is held so dear because it seems essential in order for people to do anything (at least voluntarily).[1] If one accepts Davidson's view that motivating reasons are causes, then (as Kim puts it) 'agency is possible only if mental causation is possible' (1996: 127).[2] Many kinds of mental items are supposed to be causes: beliefs, desires, sensations, emotions, the contents of beliefs and desires, and the phenomenal mental properties of sensations and beliefs (i.e. those properties such that there is 'something it is like' to experience them, if sensations and beliefs have such properties). Not only are mental states supposed to be causes (and effects), but so also are mental properties. For example, my belief that dodos are extinct has the property *having the content that dodos are extinct*, and this content-property is supposed to be efficacious (e.g. with respect to how I answer questions about dodos). Likewise my experience of eating anchovies has distinctive phenomenal properties, and these properties too are supposed to be efficacious (e.g. with respect to my continuing to eat them).

Would it be a disaster to concede the inefficacy of at least some items on this list? Chalmers (1996) and Kim (1998*b*; 2005) investigate conceding the causal inefficacy of phenomenal mental properties. But it does often seem as if the *feel* of these properties is what gives rise to judgements about them. So the phenomenal mental properties associated with anchovy flavours contribute to bringing about my belief that anchovies taste good, and my action of ordering anchovy pizza. Would it be a disaster to concede that mental *properties* are inefficacious? Perhaps not; maybe it would be enough for token mental events to be efficacious. But we begin with a sincere hope to avoid this concession.

1. Six Threats to Mental Causation

We will present six recent challenges to mental causation. Each of these arguments attempts to establish that, contrary to ordinary intuitions, some large group of mental entities

is epiphenomenal. The arguments are related to each other in ways that we will not fully explore here. Although we present them as separate threats, the many different versions of them that appear in the literature often combine elements of several together. (In particular, the counterfactual test presented in the Subtraction Argument is often employed to strengthen the Qua Problem, the Causal Exclusion Argument, and the Extrinsicness Argument.)

1.1 The Problem of Strict Laws

We begin with an argument from Davidson (1970): the Problem of Strict Laws. Davidson argued that there are no strict psychophysical laws, that is, there are no exceptionless laws that connect mental characteristics with physical characteristics. Why can there be no strict psychophysical laws, and why is that a problem? First, why is it a problem? Well, on regularity views of causation, going back to Hume, in order for one event to cause another their conjunction has to be supported (/subsumed) by exceptionless laws. In that case, mental–physical causation seems to require strict psychophysical laws, and mental–mental causation seems to require strict laws connecting some mental events with other mental events. So, without these strict laws, mental causation appears impossible. Second, why can there be no strict psychophysical laws? Davidson's main argument targets intentional mental states such as beliefs and desires, and has to do with rationality. The argument is roughly this: (1) the correct overall assignment of beliefs and desires to an agent, and of action-characterizations to certain of the agent's behaviours, will be one under which the assigned actions are by-and-large rationally appropriate to the assigned beliefs and desires; thus (2) such an overall 'intentional interpretation' of an agent is always potentially open to revision, should that become necessary in order to accommodate subsequent behaviours that might otherwise fail to qualify as rationally appropriate; but (3) such openness to potential revision is incompatible with there being strict psychophysical laws; hence (4) there are no such laws.

This argument challenged strict psychophysical laws both diachronic and synchronic. In challenging strict synchronic psychophysical laws, it thereby also challenged the views asserting that mental *types*—that is, mental properties or mental event-kinds—are identical to physical types (e.g. Feigl 1958; Smart 1959).[3]

In response, Davidson presented anomalous monism. The monism part of his view says that all events are physical: events falling under mental kinds or types are just the same as certain events falling under physical types, that is, token mental events are identical to token physical events. The anomalousness part says that there are no exceptionless laws connecting mental event-types (mental properties) to physical event-types. In other words, there are no strict psychophysical laws. This was a retreat to a psychophysical token-identity theory[4] (unlike Smart's or Feigl's psychophysical type-identity theories), but Davidson thought this retreat was necessary in order to counter the Problem of Strict Laws. Material monism avoids the threat to mental causation because there *are* strict physical–physical laws. So, in spite of the anomalousness of the mental, mental events are causes by virtue of falling under physical descriptions.

And so anomalous monism avoids the Problem of Strict Laws. It remains a problem,

however, for any views that accept (1) that mental causation requires strict psychophysical laws together with (2) the absence of strict psychophysical laws.

1.2 The Qua Problem

Because Davidson's anomalous monism is a version of token identity theory, on this view the efficacy of mental events is no more at risk than the efficacy of physical events. But the focus now moves from events to properties: it has been argued that anomalous monism doesn't allow for the efficacy of mental *properties*. Given that mental/physical event *c* is a cause of event *e*, is it in virtue of a mental property of *c* or merely in virtue of a physical property of *c* that it caused *e*? If it is not in virtue of a mental property that *c* is a cause of *e*, then in an important respect mentality has again fallen out of the picture. The problem has often been expressed with the '*qua*' locution: is *c* efficacious *qua* mental or only *qua* physical? It has also been expressed in terms of relevance: are the mental properties of the cause *causally relevant*?

The following examples show how natural it is to think of the cause of an event as being the cause *qua* some particular property it has (and not *qua* its other properties):

> A gun goes off, a shot is fired and it kills someone … The loudness of the shot has no causal relevance to the death of the victim. Had the gun been equipped with a silencer the shot would have killed the victim just the same. (Sosa 1984: 277–8)

> Meaningful sounds, if they occur at the right pitch and amplitude, can shatter glass, but the fact that these sounds have a meaning is irrelevant to their having this effect. The glass would shatter if the sounds meant something completely different or if they meant nothing at all. (Dretske 1989: 1–2)

In the first example, the shot was a cause of the death, but it was not in virtue of its loudness that it was a cause. In the second example, the sound was a cause of the shattering of the glass, but this was not in virtue of being meaningful, but in virtue of its loudness and pitch. In the same way, it would seem, on anomalous monism there is no reason to think that a physical event is caused *qua* the mental properties of its mental/physical cause.[5] The physical properties of the mental/physical cause are sufficient for the effect, and so there is no work left for the mental properties to do.

Although the Qua Problem was originally aimed at Davidson's anomalous monism, more generally one can see that it is, in principle, a problem for any view of the mental except type identity theory.[6]

1.3 The Subtraction Argument

The Subtraction Argument is similar to the Qua Problem in that the project is to find some

causal work that is done by mental properties. In the case of the Subtraction Argument, however, it is the causal efficacy of *phenomenal* mental properties that is the main focus.[7] David Chalmers, who himself is a prime target for the Subtraction Argument (because he affirms that there are metaphysically possible worlds that are physically just like ours but in which there is no phenomenal consciousness), summarizes it succinctly:

> If it is possible to subtract the phenomenal from our world and still retain a causally closed world Z, then everything that happens in Z has a causal explanation that is independent of the phenomenal … But everything that happens in Z also happens in our world, so the causal explanation that applies in Z applies equally here. So the phenomenal is causally irrelevant. (1996: 150)

On Chalmers's view, it is metaphysically possible (though not nomically possible) for there to be a causally closed world which is just like our own, but without the phenomenal, so it seems to follow that the phenomenal is causally irrelevant.[8] The argument relies on an intuitive test for causal efficacy, such as:

> Makes No Difference (MND): If it is (non-vacuously)[9] true that had the cause lacked property F then the effect would still have occurred, then property F is causally irrelevant to such an effect.

This test appeals to the intuition that if a property makes no difference to an effect then it is irrelevant to that effect. Although this test does involve a counterfactual, it does not presuppose a full-blown counterfactual analysis of causal relevance.[10] The Subtraction Argument applies to any view that says of at least some mental properties that they are not only not identical to physical properties, but that they also do not supervene in a metaphysically necessary way on physical properties.

1.4 The Causal Exclusion Argument

The Causal Exclusion Argument is much like the Subtraction Argument in that it challenges the belief that there is some causal work to be done by the mental. This argument attempts to show that mental properties are causally irrelevant on the grounds that fundamental physics already gives a complete causal story. It can be formulated thus:

(1) The Causal Closure of the Physical: every physical event that has a sufficient cause at time *t* has a sufficient physical cause at *t* (at least, sufficient disregarding quantum indeterminacy).

(2) (a) No event has more than one sufficient cause at a time, unless it is causally

overdetermined.

   (b) A physical event is not causally overdetermined by both the instantiation of a mental property and the instantiation of a physical property.[11]

(3) Mental properties are distinct from physical properties.

(4) If a physical event e has a sufficient physical cause at time t, and e is not causally overdetermined, then no non-physical property that is instantiated at t is causally relevant to the occurrence of e.

Therefore

(5) Epiphenomenalism: mental properties are never causally relevant to the occurrence of a physical event.[12]

Cases of redundant causation involve two (or more) events which are each individually sufficient for an effect but neither of which is necessary for the effect. Cases of redundant causation can be divided into cases of pre-emption, where only one of the events counts as the cause, and cases of overdetermination. So we understand 'overdetermination' to describe cases in which, at a given time, (i) the instantiation of property F_1 is individually sufficient for the effect, (ii) the instantiation of property F_2 is likewise individually sufficient, (iii) neither property's instantiation is necessary for the effect, at a given time, and furthermore (iv) there is no pre-emption.[13]

From this definition of overdetermination, premiss 2(a), above, follows readily. Claim 2(b) is thus a more substantive one—but it also looks difficult to deny, given that fundamental physics gives a complete causal story by itself. Premiss 4 expresses the core thought behind the argument: the idea that the sufficient physical cause does all the work *by itself* in generating the effect-event, thereby rendering causally irrelevant any simultaneously instantiated mental properties. The Causal Exclusion Argument threatens the causal efficacy of mental properties for every view that is not a form of type identity.

1.5 The Extrinsicness of Content Argument

The Extrinsicness of Content Argument threatens intentional mental properties, such as belief properties or desire properties. It attempts to show that intentional mental properties are epiphenomenal, particularly for those who think that intentional mental properties have wide content (where 'wide content' refers to content that is fixed, at least partially, extrinsically).

By way of illustration, consider a classic example: Hilary Putnam's Twin Earth, a place identical to our world in every respect except that the stuff referred to as 'water' on Twin Earth is actually composed of the chemical compound XYZ rather than H_2O (Putnam 1975).

By hypothesis, your duplicate on Twin Earth would be virtually an exact duplicate of you, instantiating virtually the same physical properties and displaying matching behaviour. An externalist would say that when you desire a glass of water, your duplicate's corresponding desire is importantly different from yours, because the respective desires are for different things (namely, H_2O and XYZ). It is this sort of externalist position about intentional mental properties that is challenged by the Extrinsicness of Content Argument.

Mental properties of an individual, like believing that *p*, desiring that *p*, knowing that *p*, and other propositional-attitude properties, are often held to be extrinsically individuated. In other words, the individuation conditions of such properties are often taken to depend, at least partly, on relational facts linking the individual to its environment—for example, causal facts, or historical/evolutionary facts, or facts about co-variation between internal states and external conditions. One reason for this position is that people have a strong response to the Twin Earth thought experiment. They find it plausible to think that one's belief about water refers to H_2O whereas the corresponding belief of one's twin on Twin Earth refers to XYZ.[14]

The point of the Extrinsicness of Content Argument is that extrinsically individuated factors make no difference to an effect. 'You do the same thing—drink—whether it is water (H_2O) you desire or twater (XYZ) … That extrinsic factors can be varied indefinitely at no cost to the effect exposes them as irrelevant hangers-on' (Yablo 2003: 316–17). This is very closely related to the Qua Problem and the Subtraction Argument in this way: extrinsic factors are thought to be causally irrelevant because they make no difference to the effect, and thus extrinsically individuated mental properties are thought to be likewise causally irrelevant.

Consider a simple belief-behaviour chain. You believe something and this causes you to say it is so. You believe that all water beds contain water and this causes you to say, 'All water beds contain water.' But which properties of your belief were causally relevant? In particular, was the content property of your belief causally relevant to your behaviour? That is, was the property of being a belief with the content that all water beds contain water causally relevant? It seems not, because your duplicate on Twin Earth would have a belief with different content, but would, nevertheless, make the same vocalization. The physical event, your vocalization, cannot be the effect of your (wide) belief, since changes to the content of that belief make no difference for the occurrence of the physical event.

This worry can be applied to any mental property that is both intentional and whose content is extrinsically individuated. Many in contemporary philosophy of mind hold that all mental intentionality is like that—which would mean that all intentional mental properties are subject to the Extrinsicness of Content Argument.[15]

1.6 The Dormitivity Argument: Problems with Functional Properties

The Dormitivity Argument targets functionalism, the view that mental properties are identical to certain functional properties. A functional property is a higher-order property: the property of having some lower-order property that itself has a certain specific causal role—a role characterized by a pattern of causal relations to sensory inputs, behavioural outputs, and other properties in the pattern. For example, the property of feeling hungry may be the property of having some physical property that is typically caused by not eating for some time

(an input), typically causes foraging for food (an output), and typically causes (among other things) another physical property that realizes the desire to eat (another functional property). Block (1990) motivates concern about the causal relevance of functional properties by considering the causal relevance of second-order properties in general, where a second-order property, as Block defines it, is only slightly more general than a functional property. A second-order property is the property of having some first-order property or other with certain causal relations to other first-order properties.

Suppose that you take a sleeping pill and fall asleep. Block argues that the property of dormitivity, that is, the second-order property of having some property or other that is causally relevant to sleep, is not causally relevant to your falling asleep. He says,

If a dormitive pill is slipped into your food without your noticing, the property of the pill that is causally relevant to your falling asleep is a (presumably first-order) chemical property, not, it would seem, the dormitivity itself. Different dormitive potions will act via different chemical properties, one in the case of Valium, another in the case of Seconal. But unless you know about the dormitivity of the pill, how could the dormitivity itself be causally relevant to your falling asleep? (Block 1990: 155–6)

Block has at least two separate worries about the efficacy of functional properties, both arising from the intuition that dormitivity is inefficacious. The first has to do with exclusion and overdetermination again. This gives rise to an important version of the exclusion problem, according to which it is the first-order 'filler' property that does the causal work, leaving none for the second-order 'role' property to do. The second comes from the fact that dormitivity is logically or definitionally related to sleep. We have already discussed the exclusion problem, so let us concentrate on the new worry then. Block's argument goes something like this:

(1) A second-order property is defined as the property of having some first-order property or other with certain causal relations to other first-order properties.

(2) First-order properties are distinct from second-order properties.

(3) The fact that instantiation of the second-order property is followed by states with those effects is explained in terms of the logical connection stated in its definition.

(4) Although logical relations do not preclude causal connections, when logical relations are present, an additional reason is needed in order to be justified in believing that a causal connection exists.

(5) There are no such reasons in the case of the functional properties that, according to functionalism, are identical to mental properties.

Therefore

(6) these functional properties are not causally relevant to the effects in terms of which

they are defined.

2. Proposed Solutions to the Problems

In this section, we consider a few of the prima facie plausible solutions that have been proposed to the problems we have described in sect. 1. We do not intend to give a comprehensive survey of the many interesting positions on this subject, since such a task could not be performed in a chapter of this length. Our goal in presenting these proposals is to demonstrate that (at least some of) the leading candidate solutions proposed so far can resolve only some of the problems in sect. 1—not all of them at once. It is our contention that this is common among the proposed solutions, with the exception of contextualism, and so this is a point in favour of contextualism.

2.1 Type Identity

One way of attempting to save mental causation from epiphenomenalism is via type identity. We mentioned earlier that type identity is a way to circumvent the Qua Problem. This is so because type identity does not differentiate between mental and physical properties. Because of this, it also circumvents nearly all the problems we described. Type identity runs into the Problem of Strict Laws, however, as we mentioned above. If Davidson's argument against strict psychophysical laws is successful, then it also works against the claim that mental properties are identical to physical properties (as explained in sect. 1.1).

Another very familiar worry about type identity, typically stressed by functionalists, is the problem of multiple realizability. The argument is this: there cannot be psychophysical property identities on the grounds that such identities would eliminate the possibility that a mental property could be physically realized in different ways by different beings.

The problem is not quite as bad as is often thought, however. Lewis (1966; 1980) explains how a type identity theory can avoid this problem. The idea is that mental concepts can be thought of as functionally definable non-rigid concepts that non-rigidly designate different physical properties, depending upon the kind of being in question.[16] For example, if one considers the 'pain' concept as functionally definable, this means that pain is identical to a particular sort of physical phenomenon *for a particular kind of being*, but may be identical to a different sort of physical phenomenon for some other kind of being. For example, suppose that the property that fills the pain-role in humans differs from the property that fills the pain-role in dogs. That is, the mental phenomenon of pain is realized differently in humans than it is in dogs. By relativizing the identity of mental and physical phenomena to a kind, Lewis can explain this example of multiple realization easily.[17] Kinds need not even be as broad as whole species. Lewis allows that a single being could be a kind of its own, and so multiple realization among members of the same species can also be accounted for in this way.

Still, type identity must contend with what is known as *strong* multiple realization—the occurrence of multiple realization within a single being at a single stage of its life.[18] It cannot simply be taken for granted that every instance of a particular sort of mental event within a

kind of creature is identical to an instance of some single sort of physical event, even if we allow that a single, individual being can be counted as a kind. To illustrate, consider the possibility that you experience a specific kind of pain today, and then again tomorrow, but although the mental property is the same on both occasions, the underlying physical property is different. This is surely a genuine epistemic possibility, and for all science can tell us at this time, an actuality. Indeed, it is also epistemically possible that this pain-property is physically realizable either way in you *right now,* by two different neurophysical properties. And in any case, surely it is epistemically possible there are *physically possible* creatures (and ergo, it is epistemically possible that there are *metaphysically* possible creatures) in whom mental properties are strongly multiply realizable in this way—whether or not this is so for humans. An acceptable theory of mind needs to accommodate these epistemic possibilities—which type-identity theory does not.[19]

2.2 Kim on Mental Causation[20]

Kim has been wrestling with the issue of mental causation for a long time, and so it is not surprising that he is sensitive to many of the problems we have raised. The problem of causal exclusion, in particular, has been forcefully emphasized by Kim over the years. Although his thinking on these matters has evolved, two themes have been consistently prominent in his discussions: first, an ongoing reluctance to flatly embrace a type–type psychophysical identity theory, and second, an awareness of the looming threat to mental causation posed by the exclusion problem *unless* one embraces a type–type identity theory. Here we will mention two fairly recent proposals he has tried out, by way of seeking to avoid the exclusion problem without embracing type–type psychophysical identities.[21]

The first proposal treats mental properties as second-order functional properties, and embraces a version of psychophysical token-identity theory. Here is a key passage. (Bear in mind that for Kim, a token event or state is a property-instance—that is, an entity consisting of an object's instantiating a property at a time.)

The main idea … is this: if a given instance of M occurs in virtue of being realized by P, the M-instance and its P-realizer do not compete for causal role … When we speak of causal powers of M as such, we are speaking disjunctively of the causal powers of the P_i's … [F]or M to be instantiated on a given occasion *is* for an appropriate P to be instantiated on that occasion and in an appropriate causal environment. There is no further fact of the matter, as one might say, to M's instantiation on this occasion beyond P_i's instantiation in the particular context involved … Given this general picture, a simple solution to the exclusion problem suggests itself … [A]ny given M-instance must be either a P_1-instance or … , where $P_1, P_2, …$ are realizers of M … [A]n M-instance is identical with a P_i-instance, for some M-realizer P_i, and hence there is one event here not two, and this dissipates the causal competition. (Kim 1993: 362–4)[22]

Does this approach solve the problem of causal exclusion, as Kim claims it does? We think not, as the following line of reasoning shows. Suppose a mental property M is instantiated by creature C at time t; and let P_r be the first-order physical property that realizes M on this occasion. Thus the concrete M-event is the physical event $[C, P_r, t]$, that is, the event consisting of the creature C instantiating P_r at t. Suppose that the mental event produces a subsequent physical effect—say, a subsequent physical event $[C, P^*, t+\Delta]$ consisting of C instantiating some physical property P^* at a time $t+\Delta$. The property P_r, the constitutive property of the mental cause-event, is clearly a causal property—a property whose instantiation 'does causal work' in generating the effect. But now the problem of causal exclusion arises all over again for property M. Although the proposal does render token mental events causally efficacious qua physical, considerations of causal exclusion suggest that token mental events are not causally efficacious *qua* mental—that is, that mental properties are not causal properties. If causal-exclusionary reasoning works at all, then it apparently applies to this proposal of Kim's no less than to any other version of property dualism.

A more recent approach Kim has tried, again appealing to functionalist thinking in philosophy of mind, has been to back away from the idea that there are distinct mental properties at all—and to invoke mental *concepts* that supposedly designate various different physical properties 'disjunctively'. He writes:

> We must now confront the following question: If M_i is a second-order property and P_i a first-order property, or if M_i is a causal role and P_i is the occupant of that role, how could they be one and the same thing … ? We may begin by explicitly recognizing that by existential quantification over a given set of properties, we do not literally bring into being a new set of properties … So it is less misleading to speak of second-order *descriptions* of *designators* of properties, or second-order *concepts*, than second-order properties … On the present view, the concepts introduced by second-order designators pick out first-order properties disjunctively. When I say, 'x has property M', where M is a second-order designator (or property, if you insist), 'the truth-maker' of this statement is the fact, or state of affairs, that x has P_1 or P_2 or P_3, where the Ps are the realizers of M. (The 'or' here is sentence disjunction, not predicate disjunction; it does not introduce disjunctive predicates with disjunctive properties as their semantic values.) Suppose that in this particular case, x has M by virtue of having P_2, in which case the ultimate truth-maker of 'x has M' is the fact that x has P_2. There is no further fact of the matter to the fact that x has M over and above the fact that x has P_2. (Kim 1998a: 200)

On one construal, this passage embraces eliminativism about mental properties; it says that (1) if there were any then they would be second-order properties, but (2) there aren't any. But eliminativism about mental properties appears to throw out the baby of mental realism along with the bathwater of psychophysical property-dualism. If there are no mental properties at all, then it is harder than ever to see how token mental events could be causally efficacious qua mental. Put another way, it is harder than ever to see how causal/explanatory 'because'-statements, such as 'She winced because she was thinking about Karl Rove', could ever be

true. One does not save the causal efficacy and the explanatory relevance of mental properties by denying their existence.

On another construal, the lately quoted passage embraces not eliminativism but rather the version of the type–type psychophysical identity theory espoused by David Lewis. (Kim, so interpreted, is here espousing type–type identity without flatly acknowledging that he is.) On this view, a mental concept *M* is construed as non-rigidly designating, in a particular context of mental-property attribution, a particular one among a range of physical properties P_1, P_2, … . Which property P_i is the designated one, in a context where one judges or asserts that a given creature *C* 'has property *M*', depends on which creature-kind *C* belongs to. The concept *M* non-rigidly designates the physical property P_i, whatever it is, that occupies the *M*-role *for creatures of C's kind* . But, as emphasized already in sect. 2.1, although this Lewis-style type-identity approach does indeed handle the exclusion problem, it does so at the cost of falling foul of some of the other problems discussed above—notably, the Problem of Strict Laws and the problem of accommodating strong multiple realization.

3. CONTEXTUALISM TO THE RESCUE

We have demonstrated how difficult it is for a theory simultaneously to avoid all the problems we raised in sect. 1. In what follows we will briefly present some different styles of contextualist analysis of causation. We will then show how a contextualist approach to causation can respond nicely to the six threats to mental causation presented in sect. 1. We believe this provides significant motivation for adopting a contextualist view. However, there may be other ways to respond to the whole package of these threats, and so we do not intend to build a case for contextualism solely on the claim that it is *needed* to avoid epiphenomenalism.

3.1 Introducing Contextualism

Contextualism about a particular group of statements is the view that the meaning and truth conditions of statements in that group depend on facts about the situation in which those statements occur. For example, contextualism about knowledge is the view that statements of the form '*S* knows that *P*' and '*S* does not know that *P*' depend for their meaning and truth conditions on the context. Contextualism about causation, then, is the view that statements of the form '*c* is a cause of *e*' and '*c* is not a cause of *e*' depend for their truth-value and meaning on the context.

It may not be surprising to some that with an extra contextually fixed parameter available, one can find wriggle room to silence objections. But adopting a contextualist approach is not just to add an all-purpose 'fudge factor'; as we shall demonstrate, there is strong, independent motivation for embracing contextualism. In fact, it seems fair to say that contextual accounts of causation are a *movement* in the causation literature, a literature almost oblivious to concerns in the philosophy of mind. (Philosophy of mind, in turn, has been almost oblivious to the treatment of causation in metaphysics and philosophy of science, including contextualism

in these areas).[23]

We will briefly present two main routes for motivating contextualism about causation.[24] The first is to present linguistic evidence of the context sensitivity of causal claims. The second is to argue for a tight connection between causation and counterfactuals and then to present linguistic evidence of the context sensitivity of counterfactuals.

Example 1*: Uncle Schlomo*

This example comes from Clark Glymour and Paul Holland.

> My Uncle Schlomo smoked two packs of cigarettes a day, and I am firmly convinced that smoking two packs of cigarettes a day caused him to get lung cancer. But it may not be true that in the closest possible world in which Uncle Schlomo did not smoke two packs a day, he did not contract cancer. Reflecting on Schlomo's addictive personality, and his general weakness of will, it may well be that the closest possible world in which Schlomo did not smoke two packs of cigarettes a day is a world in which he smoked three packs a day.[25]

It seems that in the context of a conversation in which the possibility of Uncle S's smoking significantly fewer than two packs a day is considered a relevant option, this sentence makes a true statement:

> Uncle S's smoking of two packs of cigarettes a day was a cause of his cancer.

However, in the context of a conversation in which the possibility of Uncle S's smoking any fewer cigarettes than two packs a day is not considered relevant or not considered to be a realistic possibility, the same sentence makes a false statement.

Note that there may be a correct answer to the question 'If Uncle S had not smoked two packs of cigarettes a day, how many cigarettes a day would he have smoked?' Perhaps the correct answer, because of his addictive personality, somehow turns out to be that he would have smoked three packs a day. But this alone does not decide the question of whether smoking two packs of cigarettes a day was a cause of his cancer. The possibility of his smoking significantly less than two packs a day still seems to be a legitimate contrast, for the purposes of making a causal claim.

Example 2*: The Exploding Gas Tank*

A similar example comes from Field.[26]

A gas tank explodes on a jet. It is a moderately expensive model of gas tank, model T_2. Suppose that if the cheaper model T_1 had been used instead, it would still have exploded. And if the far more expensive model T_3 had been used instead, it would not have exploded.

It seems that in the context of a conversation in which the expensive model T_3 is considered a relevant option, this sentence makes a true statement:

The use of a model T_2 gas tank was a cause of the explosion.

However, in the context of a conversation in which only the cheaper model T1 is considered a relevant option, the same sentence makes a false statement.

Another route to motivating contextualism about causation is via the context sensitivity of counterfactuals.[27] Even those who reject the counterfactual analysis of causation acknowledge that causation and counterfactuals are tightly connected. Hence, it is strange that counterfactuals are widely acknowledged to be context sensitive, while the context sensitivity of causal statements is widely overlooked.

The context sensitivity of counterfactuals has often been observed.[28] Here is a version of an example from Chisholm. It seems equally reasonable to assert of CM's (non-malleable) jade ring 'If CM's ring were gold, then it would be malleable' and 'If CM's ring were gold, then some gold things would not be malleable,' though perhaps not in the same breath. What changes from one assertion to the next is the willingness to hold fixed the non-malleable nature of the ring.[29]

This example illustrates a general feature of counterfactuals: their dependence on a context-sensitive factor (whether it is overall similarity of possible worlds as in Lewis's picture, or tacit assumptions as in Chisholm's picture).

Because of the tight connection between counterfactuals and causal claims, the context sensitivity of counterfactuals carries over to causal claims. This is clearly a subtle feature of causal claims, for people usually do not notice it when they make causal claims. But however subtle, we maintain that it is a significant feature: overlooking it can sometimes lead people astray. In particular, as we shall show, this can happen in assessing the threat of epiphenomenalism.

3.2 Contextualist Views of Causation

Contextualism about causation does not have a long history. Van Fraassen argued for a contextual view of explanation (not causation) in 1980, but the first explicitly contextual view of causation seems to be one discussed by Unger in 1984. Then in 1986 was Holland and Glymour, in 1989 Horgan, and in the 1990s Hitchcock and others.

It is also worth noticing that others employ context-dependent variables in their accounts of causation without advertising these accounts as contextualist—for example, Lewis (2004), Yablo (2002), and Collins (2004). Even in his original paper on causation, Lewis observes that 'The vagueness of similarity does infect causation, and no correct analysis can deny it' ([1973*a*] 1986: 163). His theory already permitted context dependence although he did not pursue that line further. The causal modelling/interventionist approach to causation (see e.g. Meek and Glymour (1994); Pearl (2000); Hitchcock (2001); Woodward (2003)) that is currently popular, especially in statistics and computer science, seems to require a contextualist interpretation, although this feature of the approach is not always stressed. Interventionist theories are developments of the manipulability approach to defining causation; while manipulability theories define causes in terms of manipulations of the effect, interventionist theories carefully define the notion of an allowable (human or non-human) manipulation as an intervention. In other words, variable *x* is causally relevant to variable *y* iff appropriate interventions in *x* lead to changes in *y*. This theory makes causation relative to a choice of model or set of variables.

It is important to stress the difference between Unger's contextualism about causation and most of the more recent contextualist approaches.[30] Unger considers how the distinction between the cause and background conditions is dependent on context, in particular on what is relevant or noteworthy to the speaker.[31] Putnam agrees, saying, '[O]ne man's (or extraterrestrial's) "background condition" can easily be another man's "cause" ' (1983: 214). The same kind of subjectivity in the distinction between causes and background conditions was observed by Mill and carefully described by White and others.[32] Most later metaphysicians set this 'problem of selection' aside, and specify that their purpose is to analyse the concept of being 'a cause'.[33]

Of course, the distinction between causes and background conditions is heavily dependent on interests. But contextualists such as Horgan and Hitchcock are making a bolder claim. They are willing to count both causes and background conditions as contributing causal factors, but insist that whether something counts as a causal factor *at all* depends on features of the context. Moreover, appealing to the context sensitivity of the distinction between causes and background conditions does not look as though it will illuminate mental causation, so we will concentrate on 'a cause' rather than 'the cause' in what follows.

A context-sensitive approach to 'a cause' that we find very useful simply allows restriction on the set of relevant possible worlds from context to context. And the examples given above of Uncle Schlomo and the gas tanks do seem to point to sets of relevant possible worlds as the contextually determined variable.[34]

A slightly different approach is the contrastive approach.[35] The basic idea is that just as there are sometimes explicitly contrastive causal claims, there are often (or perhaps always) *implicitly* contrastive causal claims. On one construal, ordinary causal claims are just elliptical for contrastive causal claims. For example, in one context in which I say 'Uncle S's smoking of two packs a day was a cause of his cancer', it is elliptical for 'Uncle S's smoking of two packs *rather than not smoking at all* was a cause of his cancer.' The causal relation appears at first sight to be a two-place relation, but that is to be fooled by our elliptical talk; really it is a three-place relation where the third place is a conversationally understood contrast. Note that it is quite natural on this view for there to be two separate contextual

parameters: one for contrasts to the cause and one for contrasts to the effect. In that case, the causal relation is really a four-place relation.[36]

The Contrastive approach and Relevant Worlds approach differ in the contextual parameter. With the Contrastive approach it is a contrast event or set of contrast events that may change from context to context, while with the Relevant Worlds approach it is a whole possible world or set of possible worlds that may change from context to context.[37] A different kind of contextual parameter is a set of events from the actual world that are *held fixed* when the cause is varied.[38] The contrast events of the Contrastive approach, on the other hand, are new events that are introduced instead of the cause.

Apart from contextual parameters, there is a way in which causal talk is loose that should be quite uncontroversial. 'The sun caused the wax to melt' is elliptical for 'something the sun did caused the wax to melt'. Although ordinary language undeniably employs a wide variety of causal relata in causal talk (facts, events, agents, physical objects, event aspects, properties, absences), most philosophers hold that one kind of entity is the primary causal relatum. Ordinary language claims involving other, less favoured, relata are usually presumed to be loose ways of speaking, and ultimately to be reducible to causal claims involving the primary causal relata.

Some may claim that all evidence for contextualist views of causation is better construed as evidence for context helping to narrow down causal relata. For example, one could say that in one context, 'Uncle S's smoking of two packs a day' specifies a contrastive relatum: 'Uncle S's smoking of two packs a day rather than not smoking at all' whereas in another context it specifies a different contrastive relatum: 'Uncle S's smoking of two packs a day rather than three packs a day'. Alternatively, one could say that more or less fine-grained events are designated: in the first context, 'Uncle S's smoking of two packs a day' designates the event consisting in Uncle S smoking *at all,* and in the second context it designates (at a stretch) the event consisting in Uncle S smoking *fewer than three packs a day*.[39] However, a view in which context determines the causal relata seems equally to be a contextualist view of causation (and may even be equivalent to other contextualist approaches), so there may be no real dispute here.

Which contextualist view would be the most suitable for the purposes of discussing mental causation? In the case of context dependence of counterfactuals, Lewis allows his 'similarity metric' to vary from context to context, but that choice of contextual parameter has frustrated many readers who cannot see how a metric of overall comparative similarity of possible worlds is determined by the context of an innocent musing about tail-less kangaroos (for example). Different contextualist accounts of causation may simply be equivalent formulations. What must be found is the most convenient factor to use in determining the mechanisms that fix the meanings of causal claims from the context. This is one thing that should guide the choice of contextualist account. Later in this chapter, we choose to use relevant possible worlds as a contextual parameter because we find it straightforward and convenient.

Of course, the details of mechanisms for how the context actually determines the contextual parameter are a vital part of any contextualist view. For example, on the Relevant Possible Worlds approach one needs to understand how various possible worlds may be rendered in or out of consideration by explicit or implicit assumptions in the context. In the case of

contextualism about knowledge, Lewis (1979; 1996) did a lot of work on how contextual parameters are fixed by the context, and others have since done further work. In the case of causation, among other mechanisms, something like the 'Rule of Salience' employed by contextualists about knowledge is in play. Possible worlds are rendered in or out of consideration by being taken seriously as alternative possibilities, or by being vetoed.[40] For example, one decides whether Uncle S smoking two packs a day is relevant to his lung cancer on the basis of whether, in context, the possibility of Uncle S smoking less than two packs a day is a serious alternative possibility.

Many worry that to say that causation is contextual is to take away from its objectivity. Does positing a contextualist account of causation suggest that causation is hopelessly subjective? No, all it suggests is that implicit parameters do govern the specific content of causal claims. There is a substantial objective core to causation, nonetheless. The contextual parameter is just an additional content-determining factor that usually goes unnoticed. Compare this to two other kinds of contextualism: contextualism about absolute terms such as 'flat' and contextualism about velocity. In the first case, it wasn't a major discovery that flat is contextual; it was always recognized that talk of flatness was loose. In the second case, it was a major discovery that velocity is relative to frame of reference. Causation falls somewhere between these two cases.[41]

3.3 How Contextualism Rescues the Mental from the Six Threats

In what follows, we explain in general terms how contextualism about causation can respond to the six threats we have described. We do not commit ourselves to one detailed contextual analysis of causation here but, of course, avoiding the arguments will ultimately require further development of a contextualist analysis of causation. The standard contextualist responses to each argument are quite similar to each other, but we will still go through each argument in turn.

3.3.1 *The Problem of Strict Laws and the Qua Problem Revisited*

The Problem of Strict Laws was the contention that there can be no mental–physical causation because there are no strict psychophysical laws and causation requires strict laws. Davidson's anomalous monism solves this problem for concrete token mental events, but he failed to solve the Qua Problem, for mental properties.

Contextualism can solve both the Problem of Strict Laws and the Qua Problem. First, consider the Problem of Strict Laws. When a contextualist account of causation is employed, some possible worlds are irrelevant in some contexts. Hence, causal relevance need not require exceptionless laws. Laws with exceptions will do so long as the exceptions occur only in possible worlds that are irrelevant in the context in question.[42] For example, consider this non-exceptionless generalization: if one wants something, and one believes of some act-type that performing such an act is easily within one's power and will procure the wanted item, then *ceteris paribus*, one will perform an act of that type. And consider this simple chain of

events: you undergo an occurrent desire for your freshly poured glass of whisky to be cold, and a simultaneous occurrent belief that there is ice in the fridge; straightaway, you go to the fridge and procure some ice for your whisky. Your belief and desire, *qua* mental, jointly cause your action, by virtue of falling under that non-exceptionless generalization. Possible worlds in which you refrain from acting that way despite having the belief and desire—say, possible worlds in which you also believe that someone you very much want to avoid is standing by the fridge—are simply irrelevant to the causal claim, provided that such possibilities are not pertinent in context. (In the present situation, you lack any such belief, or even the faintest suspicion that such a person might be near the fridge.)

The Qua Problem challenged the causal efficacy of mental properties; even if token mental events are identical to token physical events, the worry was that mental events are not efficacious qua mental, but only qua physical. According to anomalous monism, mental events are causes by virtue of falling under physical properties or event-kinds. This does not seem to give any assurance that mental events are efficacious *qua* mental.

A contextualist might respond in two complementary ways, thus: first, there is indeed a suitable generalization linking the instantiation of belief-properties and desire-properties to the subsequent instantiation of act-properties—namely, the one lately mentioned. Often a causal transaction between a combination of token mental states and an ensuing token-action is subsumable by such a generalization—for example, in the case of the ice and the whisky. In certain contexts of enquiry, such as contexts where one is seeking to know what reasons motivated the agent to go into the kitchen, the contextually most relevant subsuming generalization will be just such a psychological generalization, rather than (say) some physics-level law(s) that also happen to subsume the token events in question. Which generalization is most pertinent, in terms of making a causal claim on the basis of a nomic-subsumption relation between the token events in question, is itself a contextual parameter that is heavily dependent on the context of enquiry. In the context of our simple example, the psychological generalization is the most pertinent one, since one wants to know what psychological properties (among a space of possible ones) were actually instantiated by the agent and were actually motivating reasons. Hence the statement, 'She went to the fridge because of her desire to have ice in her whisky and her belief that there was ice in the fridge,' is true (with the word 'because' here signalling a causal connection). Non-strict generalizations involving non-physics-level properties can, and often do, ground such singular causal claims in many contexts of enquiry.

Second, it is a contextual matter which counterfactual variations in the cause-event count as relevant in assessing causal claims. Suppose mental and physical event c causes behaviour b. Conceivably, the mental features of event c or the physical features of event c could be varied. If one varies the *physical* features of c in the right way then it is true that b would not have occurred. But also if one varies the *mental* features of c in the right way then it is true that b would not have occurred. There is no one 'correct' way to vary the features of event c. Rather, this depends on which possible worlds are relevant or irrelevant in the context. So, in some contexts it is true to say that c is a cause of e not just in virtue of its physical features but also in virtue of its mental features. However, there may be other contexts in which the same claim of mental causation is false, and maybe even contexts in which all claims of mental causation are false.

For example, suppose that agent A instantiates the mental property *wanting a drink of water* at time t, and then says (at time $t + \Delta$) ΔMay I have a drink of water?' Suppose that the agent's instantiation of the desire-property (at t) is physically realized by the instantiation (at t) of neurophysical property N. Consider the causal claim, 'The agent's desire for a drink of water, qua mental, caused the agent's verbal behaviour.' In one context, the pertinent contrast-class might involve different psychological properties the agent could have instantiated, rather than ways (if any) that the agent's actual desire-property might have been physically realized otherwise than by property N. In this context, the agent's vocalizing behaviour counts as caused by the agent's desire, *qua* mental. In another context, however, the pertinent contrast-class might involve ways that the property N could have sub-served different mental properties—for example (on Twin Earth), the property *wanting a drink of twater*. Vary the supervenient mental property while keeping the realizing property N intact, and the behaviour still ensues.[43] In this context, the agent's vocalizing behaviour counts as *not* caused by the agent's instantiating the mental property *wanting a drink of water*.

It may seem that the contextualist is conceding the truth of epiphenomenalism in some contexts. Is mental causation really lost in these contexts? No. The contextualist will grant that under the implicit semantic parameters operative in these contexts, looking back and employing the same set of relevant possible worlds, the sentences used earlier to make true statements about mental causation now make false statements. However, causation does not literally disappear in these contexts. It just becomes difficult to make true *statements* about mental causation, when certain contextual parameters are in play. Claims about mental causation remain objectively true *under typical settings of implicit contextual parameters*.

3.3.2 *The Subtraction Argument Revisited*

The Subtraction Argument maintained that if the phenomenal is merely nomically supervenient on the physical rather than metaphysically supervenient, then there are metaphysically possible worlds in which the phenomenal does not occur, but the subsequent behaviour still occurs—thus allegedly showing that the phenomenal makes no causal difference. Suppose that there are two metaphysically possible worlds with the following features: world W_1, where a phenomenal property that was instantiated by a physical event in the actual world is cleanly excised from that physical event, and W_2, where the phenomenal property is replaced by a different phenomenal property. In both W_1 and W_2 the subsequent effect still occurs due to its sufficient physical cause, so on these grounds (according to the Subtraction Argument), the phenomenal allegedly must be epiphenomenal.

As an example of W_1, imagine a zombie (a being who is externally indiscernible from an ordinary human, but who lacks phenomenal consciousness) stopping her car at a stoplight. The cause of her stopping cannot be the phenomenal reddishness instantiated by her experience of seeing a red light at the top of the stoplight, since she has no phenomenally reddish experience. When you stop at a red light, though, it seems that you do so because of the reddish phenomenal character of your phenomenal experience as you look at the light. The possibility of zombies would indicate that although phenomenal mental states sometimes

seem like causes of physical events, they actually make no difference. As an example of W_2, one can consider the same sort of example, but rather than a zombie driver, imagine a driver whose phenomenal experiences are systematically different from your own. Perhaps the driver sees red as green and vice versa. This possibility would indicate that no particular phenomenal experience is essential for the occurrence of any physical effect.

A contextualist who grants that there are metaphysically possible worlds such as W_1 and W_2 will probably concede that there are special contexts in which one or both of these worlds are relevant because they are being taken seriously within a philosophical discussion. However, the contextualist will also argue that worlds W_1 and W_2 are irrelevant in most contexts. After all, although they are metaphysically possible, they are nomically impossible.[44] When are such nomically impossible worlds relevant, then? The contextualist will say that none are relevant in typical contexts, hence the Subtraction Argument tells us nothing here.[45]

Many philosophers have tried to block this move by claiming that because W_1 and W_2 are metaphysically possible, they are close enough to our world to be counted as relevant. We think that this is doubtful, however. Consider the nomic connection between some physical and phenomenal event-types, such as the connection between burning one's finger and experiencing pain. There are, of course, some circumstances under which this connection does not hold—you might have nerve damage in your finger that prevents you from feeling when you have burned it, for instance—but this can be accommodated with a *ceteris paribus* clause. So long as you are an ordinary person, in ordinary circumstances, burning your finger will result in your experiencing pain. What is interesting about this connection is how important it is for the actual world's being as it is. Evolutionary processes have selected for neural processes that signal specific forms of physical damage to the body and trigger appropriate responses, and such neural processes nomically realize specific kinds of pain-experience. So the resulting diachronic nomic connection between bodily-damage events and pain experiences cannot be written off as an accident. Without this diachronic connection, the world would probably be quite different than it is—humans might be far fewer in number, and would certainly have very different interests, not to mention all the other species that would be affected. Because of the *significance* of this diachronic nomic connection, we think it is reasonable to say that worlds where it does not hold are too distant to matter in most contexts.[46] And the same is true for many other nomic connections. Because contextualists can take violations of nomic connections as grounds for the irrelevance of a possible world, they are able to preserve the efficacy of the mental even if the mental supervenes on the physical only with nomic necessity, not with meta-physical necessity.

3.3.3 *The Causal Exclusion Argument Revisited*

The contextualist reply to the Causal Exclusion Argument is very much along the lines of her replies to the Qua Problem and to the Subtraction Argument. Very often, when the instantiation of a mental property M (by an agent A at a time t) is subserved by the instantiation (by A at t) of an M-realizing neural property N, a *very substantial* variation from actuality is needed to obtain a possible world in which N is instantiated by A (or by an A-like agent) but M is not simultaneously instantiated. In many contexts, the implicit parameters

governing the notion of causation will work in such a way that no such worlds are relevant.

For instance, consider a possible world W in which (1) the agent's 'causally upstream' neural interconnections (including the interconnections to sensory neurons) are so different from A's actual neural connections that the instantiation of the neural event N has none of the typical causes that N-instantiations by agent A have in the actual world, but (2) the agent's 'causally downstream' neural interconnections (including the interconnections to motor neurons) are largely the same as in actuality. Arguably, in world W the neural property N does not physically realize the mental property M—since the upstream causal role of N-instantiations is so vastly different in W than it is in the actual world. And presumably, N-instantiations in W have the same kinds of behavioural effects in W as they do in the actual world (since the agent's causally downstream neural interconnections are so similar to those of agent A in the actual world). So W is a world in which (a) N-instantiations occur in the agent without the simultaneous occurrence of M-instantiations, and (b) N-instantiations nonetheless have the same kinds of behavioural effects they have in the actual world. But the contextualist will claim that W is so utterly remote from actuality, with respect to the overall causal role of N-instantiations in the envisioned agent, that goings-on in W are simply not relevant (under typical settings of implicit contextual parameters) to claims about the causal efficacy of actual M-instantiations by the actual agent A.

As far as diagnosing the flaws in the Causal Exclusion Argument is concerned, one contextualist approach is to say that the irrelevance of such remote possible worlds as W shows the mental to be efficacious without needing to be a constitutive part of a sufficient cause. That is to say that the fourth premiss of the Causal Exclusion Argument is simply false. (Recall that the fourth premiss is: If a physical event e has a sufficient physical cause at time t, and e is not causally overdeter-mined, then no non-physical property that is instantiated at t is causally relevant to the occurrence of e.) A contextualist might also argue that the notion of 'sufficient' (as in 'each overdetermining cause is individually sufficient for the effect') is itself a contextual notion. After all, whether a cause is sufficient depends upon counter-factuals, and counterfactuals are context-sensitive.[47]

It may seem as though the contextualist is ignoring the real thrust of this argument, or as if she is just emphatically asserting that the mental and physical do not compete without explaining why. The main intuition behind the Causal Exclusion Argument seems to be that once all the causal work has been done there is no more causal work left to be done. Has the contextualist conjured up extra work for the mental to do? Well, in a sense, yes. 'Doing work' is a matter of making a difference, across a pertinent range of possible worlds. Different possible worlds count as contextually pertinent, depending (for instance) on what implicit contrast-class one has in mind in a given context. Making a difference is a contextually variable notion.

To demand a single uniform answer to the question of whether the mental makes a difference is like asking what time it is on Earth. What time it is depends (often implicitly) on what time-zone one is in. Similarly, making a difference depends on what the pertinent class of possible worlds is, according to the implicit contextual parameters governing one's current use of the notion of causation.

3.3.4 *The Extrinsicness Argument Revisited*

This argument concludes that extrinsically individuated mental properties are causally irrelevant on the grounds that the extrinsic factors that figure in their individuation make no difference to the effect. Had my desire to drink some water lacked the content property of my drinking some *water* (say, had I been a citizen of Twin Earth), my drinking behaviour still would have occurred. How would a contextualist about causation approach this argument? Quite simply, the contextualist may deem Twin Earthly worlds too distant from actuality to be relevant, under the implicit contextual parameters that normally govern mental-causation claims, to the truth value of such claims. (See Horgan 1989.)

The details of the mechanisms for determining the contextual parameter are important, but they can wait for another occasion. All we need to observe here is simply that in normal contexts, Twin-Earth scenarios do not count as relevant. (Interestingly, perhaps in other contexts one makes these possibilities relevant simply by taking them seriously.)[48]

3.3.5 *The Dormitivity Argument Revisited*

Block offered two arguments against the causal efficacy of the second-order properties that functionalists claim are identical to mental properties. The first, a version of the Exclusion Problem, was that the 'filler' property does all the causal work, leaving none for the second-order 'role' property to do. The second was that functional properties are logically related to the effects they explain in a way that allegedly precludes these properties from being causally relevant to those effects.

As an example, consider the following pharmacological case, parallel to Block's dormitive pills: Suppose that a friend described to you his drinking something and then getting very sick. You might ask whether the sickness was caused by poison. If so, then it was caused by a property of some particular chemical, namely, the second-order property of *poisonousness*. That *poisonousness* is merely a second-order property of the particular chemical ingested makes it no less reasonable for you to suspect it as the cause of your friend's sickness. Under different circumstances, though, the specific kind of poison ingested would be relevant. An emergency room doctor could not treat patients adequately if she were content with the explanation that a patient had ingested something poisonous.

This highlights the important role that contrast classes play for contextualists. When the drink in question is compared to all drinks that can make someone sick, it is reasonable to ask whether the cause of this particular sickness is the poisonousness of the drink, rather than, say, an allergic reaction in the drinker, or the drinker's intense dislike of the drink's flavour. When these are the relevant items in the contrast class, *poisonousness* is causally efficacious. If the contrast class included, instead, only poisons (e.g. arsenic, carbolic acid, and strychnine), as it might in the emergency room if the presence of poison is not in doubt and the contextually pressing need is to determine which kind of poison it is, then the drink's *poisonousness* is not a sufficiently specific property to qualify as a cause of the sickness.

Now we turn to Block's two worries, beginning with his contention that the first-order property excludes the second-order property from being causally efficacious. In order to show

that the second-order property of dormitivity is not efficacious, one would need to show that in all relevant possible worlds, whether a pill lacks the property of dormitivity makes no difference to its subsequent effect. Suppose that Block swallows a Valium pill. Subtracting the property of dormitivity, while leaving intact the chemical features definitive of Valium, requires a very large departure from actuality—for example, to a possible world in which Block's physical constitution is sufficiently different from his actual constitution that Valium has no sleep-inducing tendencies in him. The contextualist will say that in normal contexts in which a pill's dormitivity is cited as the cause of Block going to sleep, the implicit contextual parameters will rule out such a possible world as irrelevant.

Recall Block's second worry: the logical/conceptual connection between dormitivity and sleep already accounts for dormitivity being followed by sleep. So, although a logical connection does not preclude causation, if a logical connection is present, some special reason is needed for postulating a causal connection. According to the contextualist, whether or not one has reason to think there is also a causal connection depends on relevant contrast classes. If one's principal interest, in asking about the pill in relation to Block's having gone to sleep, is to know whether the sleep was caused by ingesting that pill *as opposed to other possible causes besides ingesting it*—for example, by him having just consumed a bottle of wine, or by physical tiredness together with the brain's normal sleep-inducing mechanisms—then in such a context of enquiry, the pill's dormitivity does indeed count as causally relevant. But in a context of enquiry (e.g. an emergency-room scenario) in which it is already known that ingesting the pill induced sleep and the question at hand is which specific sleep-inducing chemicals were in that pill, dormitivity will not count as causally relevant.

In conclusion, we have argued here that contextualism can simultaneously address all six of the arguments for epiphenomenalism described in this chapter. And it does so in a plausible and unified way. Furthermore, we have called attention (in sect. 2) to some of the ways in which other theories cannot do this. Type identity avoids the Qua Problem, but not the Problem of Strict Laws. Kim has tried several approaches, but in so far as they avoid type identity they either fail to solve the causal exclusion problem or lapse into irrealism about mental properties. Because prominent non-contextualist approaches fail to handle successfully the full package of problems, and because of the independent plausibility of contextualism about causation, contextualism should be explored more carefully as a way to avoid epiphenomenalism. It already has an important place in current thought on causation generally, but has hardly begun to be considered in the literature on mental causation. We contend that problems with mental causation can all be clarified and ultimately avoided by recognizing one important feature of the nature of causal claims: their dependence on context. We have only sketched responses to some of these problems. More details await a completed contextual account of causation.

Further Reading

The Problem of Strict Laws and the Qua Problem originated as attacks on Davidson (1970). Some classic responses are included in Heil and Mele (1995). Chalmers is a prime target of the Subtraction argument (also known as the 'Supervenience argument') and he sets out the argument and a response very clearly (1996: ch. 4 sect. 4). A central reading on the

Causal Exclusion argument and the Dormitivity argument is Kim (1998*b*).

   Horgan (1989) discusses some of these arguments as well as providing one of the earliest contextualist responses to them. There are only a few other discussions of the application of contextualism to mental causation: see Horgan (2001*a*), Menzies (2003), and Maslen (2005). For presentations of contextualist views of causation in general (not applied to the problems of mental causation), Hitchcock (1996), Maslen (2004*a*), Menzies (2007), and Schaffer (2005) are recommended as accessible and persuasive.

REFERENCES

BENNETT, K. (2003). 'Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It', *Noûs* 37/3: 471–97.

BLOCK, N. (1990). 'Can the Mind Change the World?', in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam.* Cambridge: Cambridge University Press, 137–70.

CARROLL, J. (2003). 'Making Exclusion Matter Less'. Manuscript.

CHALMERS, D. (1996). *The Conscious Mind.* Oxford: Oxford University Press.

CHISHOLM, RM. (1955). 'Law Statements and Counterfactual Inference', *Analysis* 15: 97–105.

COLLINS, J. (2004). 'Preemptive Prevention', in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals.* Cambridge, Mass.: MIT, 107–18.

DAVIDSON, DONALD (1970). 'Mental Events', in Lawrence Foster and J. W. Swanson (eds.), *Experience and Theory.* London: Duckworth; repr. in his (2001) *Essays on Actions and Events.* Oxford: Clarendon, 2nd edn.

DRETSKE, F. (1989). 'Reasons and Causes', *Philosophical Perspectives* 3: 1–15.

FEIGL, H. (1958). 'The "Mental" and the "Physical" ', in H. Feigl, M. Scriven, and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science.* Minneapolis: University of Minnesota Press, ii.

FIELD, H. (1997). Student Seminar on Causation at University of New York. Unpublished.

GOODMAN, N. (1983). *Fact, Fiction, and Forecast.* Cambridge, Mass.: Harvard University Press.

HEIL, J., and MELE, A. (1995). *Mental Causation.* Oxford: Clarendon.

HITCHCOCK, C. R. (1996). 'Farewell to Binary Causation', *Canadian Journal of Philosophy* 26/2: 267–82.

—— (2001). 'The Intransitivity of Causation Revealed in equations and Graphs', *Journal of Philosophy* 98: 273–99.

HOLLAND, P. (1986). 'Statistics and Causal Inference', *Journal of the American Statistical Association* 81: 945–60.

HORGAN, T. (1980). 'Humean Causation and Kim's Theory of Events', *Canadian Journal of Philosophy* 10: 663–79.

—— (1987). 'Supervenient Qualia', *Philosophical Review* 96/4: 491–520.

—— (1989). 'Mental Quausation', *Philosophical Perspectives* 3: 47–76.

—— (1998). 'Kim on Mental Causation and Causal Exclusion', *Philosophical*

Perspectives 11: 165–84.

—— (2001*a*.) 'Causal Compatibilism and the Exclusion Problem', *Theoria—Segunda Época* 16/1: 95–116.

—— (2001*b*). 'Multiple Reference, Multiple Realization, and the Reduction of Mind', in F. Siebelt and B. Preyer (eds.), *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*. Oxford: Rowman & Littlefield.

—— (2007). 'Mental Causation and the Agent-Exclusion Problem', *Erkenntnis 67/2: 183*–200.

—— and TIENSON, J. (1990). 'Soft Laws', *Midwest Studies in Philosophy* 15: 256–79.

— — (2002). 'The Intentionality of Phenomenology and the Phenomenology of Intentionality', in D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press.

— — and GRAHAM, G. (2004). 'Phenomenal Intentionality and the Brain in a Vat', in R. Schantz (ed.), *The Externalist Challenge*. New York: Walter de Gruyter, 297–317.

KIM, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge: Cambridge University Press.

—— (1996). *Philosophy of Mind*. Dimensions of Philosophy Series. Boulder: Westview.

—— (1998*a*). 'The Mind–Body Problem: Taking Stock after Forty Years', *Philosophical Perspectives* 11: 185–207.

—— (1998*b*). *Mind in a Physical World*. Cambridge, Mass.: MIT.

—— (2005). *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

LE PORE, E., and LOEWER, B. (1987). 'Mind Matters', *Journal of Philosophy* 84/11: 630–42.

LEWIS, D. (1966). 'An Argument for the Identity Theory', *Journal of Philosophy* 63: 17–25.

—— ([1973*a*] 1986). 'Causation', *Journal of Philosophy* 70: 556–67; repr. in Lewis, *Philosophical Papers II*. Oxford: Oxford University Press.

—— (1973*b*). *Counterfactuals*. Cambridge, Mass: Harvard University Press.

—— (1973*c*). 'Counterfactuals and Comparative Possibility', *Journal of Philosophical Logic* 2: 418–46.

—— (1979). 'Scorekeeping in a Language Game', *Journal of Philosophical Logic* 8: 339–59.

—— (1980). 'Mad Pain and Martian Pain', in Ned Block (ed.), *Readings in Philosophy of Psychology*. Cambridge, Mass.: Harvard University Press.

—— (1996). 'Elusive Knowledge', *Australasian Journal of Philosophy* 74/4: 549–67.

—— (2004). 'Causation as Influence', in John Collins, Ned Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 75–106.

LOAR, B. (2003). 'Phenomenal Intentionality as the Basis of Mental Content', in M. Hahn and B. Ramberg (eds.), *Reflections and Replies*. Cambridge, Mass.: MIT.

McGINN, C. (1989). *Mental Content*. Oxford: Blackwell.

—— (1991). *The Problem of Consciousness: Essays towards a Resolution*. Oxford: Blackwell.

McLAUGHLIN, B. (1989). 'Type Epiphenomenalism, Type Dualism, and the Causal

Priority of the Physical', *Philosophical Perspectives* 3: 109–35.

MASLEN, C. (2004*a*). 'Causes, Contrasts and the Nontransitivity of Causation', in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 341–58.

—— (2004*b*). 'Elusive Causation'. Manuscript.

—— (2005). 'A New Cure for Epiphobia: A Context-Sensitive Account of Causal Relevance', *Southern Journal of Philosophy* 43/1: 131–46.

MEEK C., and GLYMOUR, C. (1994). 'Conditioning and Intervening', *British Journal for the Philosophy of Science* 45: 1001–21.

MENZIES, P. (2003). 'The Causal Efficacy of Mental States', in S. Walter and H. D. Heckmann (eds.), *Physicalism and Mental Causation: The Metaphysics of Mind and Action*. Exeter: Imprint Academic.

—— (2004). 'Difference-Making in Context', in John Collins, Ned Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 139–80.

—— (2006). Book Review of *Making Things Happen: A Theory of Causal Explanation. Mind* 115: 821–6.

—— (2007). 'Causation in Context', in Huw Price and Richard Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press, 191–223.

—— (2008). 'The Exclusion Problem, the Determination Relation, and Contrastive Causation', in J. Hohwy and J. Kallestrup (eds.), *Being Reduced—New Essays on Reduction, Explanation, and Causation*. Oxford: Oxford University Press.

MILL, J. S. (1872). *A System of Logic, Ratiocinative and Inductive*. London: Longmans, Green, Reader & Dyer.

PAUL, L. A. (2004). 'Aspect Causation', in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 205–24.

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

PRICE, H. (2007). 'Causal Perspectivalism', in H. Price and R. Corry (eds.), *Causation, Physics and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press, 250–92.

PUTNAM, H. (1975). 'The Meaning of "Meaning"', *Mind, Language and Reality: Philosophical Papers*. Cambridge: Cambridge University Press, ii.

—— (1983). 'Why There Isn't a Ready-Made World', *Realism and Reason*. Cambridge: Cambridge University Press.

SCHAFFER, J. (2005). 'Contrastive Causation', *Philosophical Review* 114: 297–328.

SIEWERT, C. (1998). *The Significance of Consciousness*. Princeton: Princeton University Press.

SMART, J. J. C. (1959). 'Sensations and Brain Processes', *Philosophical Review* 68/2: 141–56.

SOSA, E. (1984). 'Mind–Body Interaction and Supervenient Causation', *Midwest Studies in Philosophy* 9: 271–82.

STALNAKER, R. (1968). 'A Theory of Conditionals', in N. Rescher (ed.), *Studies in Logical Theory*. American Philosophical Quarterly Monograph 2. Oxford: Blackwell.

STICH, S., and WARFIELD, T. (eds.) (1994). *Mental Representation: A Reader*. Oxford:

Blackwell.

UNGER, P. (1984). *Philosophical Relativity*. Oxford: Oxford University Press.

VAN FRAASSEN, B. C. (1980). *The Scientific Image*. Oxford: Clarendon.

WOODWARD, J. (2003). *Making Things Happen*. New York: Oxford University Press.

WHITE, M. (1965). *Foundations of Historical Knowledge*. New York: Harper & Row.

YABLO, S. (1997). 'Wide Causation', *Mind, Causation and the World: Philosophical Perspectives* 11: 225–81.

—— (2002). 'De Facto Dependence', *Journal of Philosophy* 99: 130–48.

—— (2003). 'Causal Relevance', *Philosophical Issues* 13: 316–29.

—— (2004). 'Advertisement for a Sketch of an Outline of a Prototheory of Causation', in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 119–38.

CHAPTER 25
CAUSATION, ACTION, AND FREE WILL

ALFRED R. MELE

MANY issues at the heart of the philosophy of action and of philosophical work on free will are framed partly in terms of causation. The leading approach to understanding both the nature of action and the explanation or production of actions emphasizes causation. What may be termed *standard causalism* is the conjunction of the following two theses: (1) an event's being an action depends on how it was caused; (2) proper explanations of actions are causal explanations.[1] Important questions debated in the literature on free will include: is an action's being deterministically caused incompatible with its being freely performed? Are actions free only if they are indeterministically caused? Does the indeterministic causation of an action preclude its being freely performed? Does free action require agent causation?

In this chapter, I concentrate on issues about action and free will that centrally involve causation. The topics of sects. 1 and 2 are causalism about action and two alleged problems for it. The main business of sects. 3 and 4 is a critical survey of some leading positions on the production of free actions.

1. CAUSALISM ABOUT ACTION

Typical causal theories of action feature as causes such mental items as beliefs, desires, intentions, and related events (e.g. acquiring an intention). An attractive theory of action treats actions as being analogous to money in an important respect. The coin with which I just purchased a snack is a genuine US dollar partly in virtue of its having been produced (in the right way) by the US Treasury Department. A duplicate coin produced with plates and metal stolen from the Treasury Department is a counterfeit US dollar, not a genuine one. Similarly, according to one kind of causal theory of action, a certain event is my buying a snack—an action—partly in virtue of its having been produced 'in the right way' by certain mental items (Davidson 1980; Brand 1984).[2] An event someone else covertly produces by remote control—one including visually indistinguishable bodily motions not appropriately produced by mental states of mine—is not a buying of a snack by me, even if I feel as though I am in charge.[3]

Causalism is normally associated with a naturalistic stand on agency according to which mental items featured in causal explanations of the actions of physical beings in some way depend on or are realized by physical states and events. In principle, causalists can welcome any viable solution to the mind–body problem that supports an important place for 'the mental' in causal explanations of actions, including viable solutions according to which what

does the causal work (in physical agents) are physical states and events that realize beliefs, desires, intentions, events of intention acquisition, and the like. Arguably, a mental item that figures in a genuine causal explanation of an action need not itself *be* a cause; its place in such an explanation may be secured partly by its relation to a physical cause that realizes it (see Jackson 2000; Jackson and Pettit 1988; 1990; Mele 1992: ch. 2). Causalism also is non-restrictive about free will. Although some causalists endorse compatibilism (the thesis that free action is compatible with determinism), that is optional for them. Provided that causation is not essentially deterministic, causalists can embrace libertarianism, the conjunction of incompatibilism and the thesis that there are free actions. Some non-causalists are incompatibilists, but there is no entailment here. Harry Frankfurt, a compatibilist (1988: chs. 1–2), rejects causalism (ch. 6).

The idea that actions are to be explained, causally, in terms of mental states or events is at least as old as Aristotle: 'the origin of action—its efficient, not its final cause—is choice, and that of choice is desire and reasoning with a view to an end' (1915: 1139^{a}31–2). It continues to have a considerable following (including Bishop 1989; Brand 1984; Davidson 1980; Goldman 1970; Mele 1992; 2003; Velleman 1989, 2000). Owing partly to the influence of Wittgenstein (1953) and Ryle (1949), causalism fell into disfavour for a time. The first major source of its revival is Donald Davidson's 'Actions, Reasons, and Causes' (1963; 1980: ch. 1).

There, in addition to rebutting familiar arguments against causalism and developing a positive causalist view, Davidson presents non-causalists with a difficult challenge. Addressed to philosophers who hold that when we act intentionally we act for reasons, the challenge is to provide an account of the reasons *for which* we in fact act that does not treat (our having) those reasons as figuring in the causation of the relevant behaviour (or, one might add, as realized in physical causes of the behaviour). The challenge is especially acute when an agent has two or more reasons for *A*-ing but *A*-s for only one of them.[4] Here is an illustration:

Al has a pair of reasons for mowing his lawn this morning. First, he wants to mow it this week and he believes that this morning is the most convenient time. Second, Al has an urge to repay his neighbour for the rude awakening he suffered recently when she turned on her mower at the crack of dawn and he believes that his mowing his lawn this morning would constitute suitable repayment. As it happens, Al mows his lawn this morning only for one of these reasons. In virtue of what is it true that he mowed his lawn for this reason, and not the other, if not that this reason (or his having it), and not the other, played a suitable causal role in his mowing his lawn? (Mele 1997*a*: 240)

In Mele (2003: ch. 2), I review detailed attempts to answer this challenge (Ginet 1990; Sehon 1994; Wallace 1999; Wilson 1989) and argue that they fail. Space constraints preclude pursuing the issue in any detail here, but I will say a bit more about it.

Philosophers who hold that reasons are causes of actions tend to regard reasons as states of mind and, more specifically, as belief-desire pairs (e.g. the pair constituted by Al's desire to avenge his rude awakening and his belief about how to do that). Whether states of mind are actually *reasons* for action is a disputed issue, but the dispute is incidental to the matter at

hand.[5] If, for example, only propositions count as reasons for action, a causalist will say that a physical agent acts for a reason *p* (where *p* is a proposition) only if the agent's belief that *p* (or the belief's physical realizer) plays a causal role in the production of the action. One can say, in a neutral way, that Al has a 'convenience' reason (*R*1) and a 'vengeance' reason (*R*2) for mowing his lawn. Assuming a familiar kind of naturalism and physical agents (as opposed e.g. to immaterial spirits), if reasons are states of mind, reasons have physical realizers, and if reasons are something else, agents' attitudes towards them have physical realizers. Consider now the physical realizers associated with *R*1 and *R*2: call them, respectively, *rR*1 and *rR*2. Imagine that Al is so constituted that a neuroscientist can, without altering *rR*1 itself, render it incapable of having any effect on Al's bodily motions while allowing *rR*2 to figure normally in the production of bodily motions. Suppose that a neuroscientist does this to Al before he starts mowing his lawn and that *rR*1 is impotent in the imagined way the entire time he is mowing. Is it nevertheless possible that Al mowed his lawn for reason *R*1, and, if so, why? I leave the task of answering this question as an exercise for the reader.

2. Two Alleged Problems for Causalism

In this section, I discuss a pair of alleged problems for causalism: causal deviance and vanishing agents.

2.1 Causal Deviance

Deviant causal chains raise difficulties for causal analyses of action and of intentional action. The alleged problem is that whatever causes are deemed both necessary and sufficient for a resultant event's being an action or for an action's being intentional, cases can be described in which, owing to a deviant causal connection between the favoured causes (e.g. events of intention acquisition) and a resultant event, that event is not an action, or a pertinent resultant action is not done intentionally.

The most common examples of deviance divide into two types.[6] Cases of *primary deviance* raise a problem about a relatively direct causal connection between mental items (or their physical realizers) and resultant bodily motions. Cases of *secondary deviance* focus on behavioural consequences of intentional actions and on the connection between these actions and their consequences. Here are representative instances of the two types of case:

A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want [or, one might suppose, his *intention* to let go of the rope] might so unnerve him as to cause him to loosen his hold [unintentionally]. (Davidson 1980: 79)

A man may try to kill someone by shooting at him. Suppose the killer misses his victim by

a mile, but the shot stampedes a herd of wild pigs that trample the intended victim to death. (ibid. 78)

Instructive attempts to resolve the problems that such cases pose highlight four claims (Audi 1997; Brand 1984: ch.1; Harman 1997; Mele 2003: ch. 2; Mele and Moser 1997; Searle 1983: chs. 3–4; Thalberg 1984).

1. Obviously, *A* is an intentional action only if it is an action; and in many cases of deviance the pertinent event seems not to be an action. For example, the climber's 'loosening his hold' is more accurately described as the rope's slipping from his fingers. A non-action cannot falsify a claim of the form 'an action is intentional only if it is caused in way *w*'.

2. A successful analysis of intentional action may entail that there is no gap between an intentional action's psychological causal initiator (or what realizes it) and the beginning of the action. If, for example, every intentional action necessarily has the acquisition of a proximal intention (in the basic case, an intention to *A straightaway*) as a *proximate* cause, there is no room between cause and the beginning of action for primary deviance.[7] Perhaps the climber's loosening his hold also is not an intentional action because it lacks a proximate cause of the right sort.

3. Intention has a continuous *guiding* function in the development of intentional action. Arguably, because the climber's loosening his hold is not guided by a pertinent intention, it is not an intentional action.

4. An action's being intentional depends on it fitting the agent's conception or representation of the manner in which it will be performed—a condition violated in Davidson's shooting scenario, standardly interpreted. (How close the fit must be obviously requires attention; see Mele and Moser (1997) on this and on ineliminable vagueness).

An analysis of intentional action that incorporates these four claims would yield the correct judgement about Davidson's two cases.

Some causal theorists who treat cases of primary deviance as attempted counterexamples to a causal account of what it is for an action to be intentional dismiss them on the grounds that they are not cases of action at all (Brand 1984: 18; Thalberg 1984). If this diagnosis is correct, primary deviance poses an apparent problem for the project of constructing a causal analysis of action. Can causalists identify something of a causal nature in virtue of which it is false that the climber performed the action of loosening his grip on the rope?

In an examination of primary deviance, Alvin Goldman remarks: 'A complete explanation of how wants and beliefs lead to intentional acts would require extensive neurophysiological information, and I do not think it is fair to demand of a philosophical analysis that it provide this information … . [A] detailed delineation of the causal process that is characteristic of intentional action is a problem mainly for the special sciences' (1970: 62, also see 166–9). Goldman's remark strikes some philosophers as evasive (Bishop 1989: 143–4; McCann 1974: 462–3), but he has a point. A deviant causal connection between an *x* and a *y* is deviant relative to normal causal routes from *x*-s to *y*-s, and what counts as *normal* here is perspective-relative. From the point of view of physics, for example, there is nothing abnormal about Davidson's

examples of deviance. And, for beings of a particular kind, the normal route from intention to action might be best articulated partly in neurophysiological terms.

One way around the problem posed by our neuroscientific ignorance is to design (in imagination, of course) an agent's motor control system. Knowing the biological being's design, we have a partial basis for distinguishing causal chains associated with overt action (that is, action essentially involving peripheral bodily motion) from deviant motion-producing chains. If we can distinguish deviant from non-deviant causal chains in agents we design—that is, chains not appropriate to action from action-producing chains—then we might be able to do the same for normal human beings if we were to know way more than we do about the human body. I pursue this line of thought in Mele (2003: ch. 2), where I develop an account of the place of intentions (or their physical realizers) in the causal initiating, sustaining, and guiding of overt actions performed by agents I design. Space constraints oblige me to set this topic aside here.

2.2 Vanishing Agents

Some philosophers contend that causalism is inconsistent with there being any actions at all, that it makes agents *vanish*. A. I. Melden (1961: 128–9) writes: 'It is futile to attempt to explain conduct through the causal efficacy of desire—all *that* can explain is further happenings, not actions performed by agents … . There is no place in this picture … even for the conduct that was to have been explained.' Thomas Nagel (1986: 110–11) expresses a similar worry:

The essential source of the problem is a view of persons and their actions as part of the order of nature, causally determined or not. That conception, if pressed, leads to the feeling that we are not agents at all … . *my doing* of an act—or the doing of an act by someone else—seems to disappear when we think of the world objectively. There seems no room for agency in [such] a world … there is only what happens.

On a straightforward interpretation, Nagel's worry is not very worrisome. Kangaroos and wombats are part of the natural order, and it seems clear that such animals act. Kangaroos and wombats fight, eat, and so on. When they do these things they are acting. The same is true of us, even if we are part of the natural order. And what is it *not* to be part of the natural order? Supernatural beings—for example, ghosts and gods—are outside the natural order. That a being needs to be supernatural in order to act is an interesting thought, but one may be excused for putting it on hold until one sees an argument for it that commands respect. Another way not to be part of the natural order is to be abstract in the way numbers are said to be. Human beings are not abstract in that way.

David Velleman voices a variant of Nagel's worry. He contends that standard causal accounts of action and its explanation do not capture what 'distinguishes human action from other animal behaviour' and do not accommodate 'human action *par excellence*' (1992: 462;see 2000:ch. 1). He also reports that his objection to what he calls 'the standard story of

human action' (1992: 461), a causal story, 'is not that it mentions mental occurrences in the agent instead of the agent himself [but] that the occurrences it mentions in the agent are no more than occurrences in him, because their involvement in an action does not add up to the agent's being involved' (ibid. 463). Velleman says that this problem would remain even if the mind–body problem were solved (ibid. 468–9), and, like Nagel (1986: 110–11), he regards the problem as 'distinct from the problem of free-will' (Velleman 1992: 465 n. 13).

Velleman here runs together two separate issues. Human agents may be involved in some of their intentional actions in ways that kangaroos and wombats are involved in many of their intentional actions. Human agents do not vanish in such actions. Scenarios in which human agents vanish are one thing; scenarios is which actions of human agents do not come up to the level of human action par excellence, whatever that may be, are another.

Typical causalists can fairly complain that Velleman has been unfair to them. In his description of 'the standard story' (1992: 461), he apparently has in mind the sort of thing found in the work of causalists searching for what is common to all (overt) intentional actions, or all (overt) actions done for reasons, and for what distinguishes actions of these broad kinds from everything else. If some non-human animals act intentionally and for reasons, a story with this topic *should* apply to them. Also, human action par excellence may be intentional action, or action done for a reason, in virtue of its having the properties identified in a standard causal analysis of these things. That the analysis does not provide sufficient conditions for, or a story about, human action par excellence is not a flaw in the analysis, given its target. If Velleman were to believe that causalism lacks the resources for accommodating human action par excellence, he might attack 'the standard story' on that front, arguing that it cannot be extended to handle such action. But Velleman himself is a causalist. Moreover, causalists have offered accounts of kinds of action—for example, free or autonomous action and action exhibiting self-control (the contrary of weakness of will)—that exceed minimal requirements for intentional action or action done for a reason.[8] Their story about minimally sufficient conditions for action of the latter kinds is not their entire story about human actions.[9]

3. FREE WILL AND LUCK

Free will may be defined as the power to act freely. But what is it to act freely? Familiar philosophical answers fall into two groups: compatibilist and incompatibilist. Compatibilism and incompatibilism are theses about the relationship between free action and determinism. *Determinism* is the thesis that a complete statement of a universe's natural laws together with a complete description of the condition of the entire universe at any point in time logically entails a complete description of the condition of the entire universe at any other point in time.[10] *Compatibilism* is the thesis that free action is compatible with the truth of determinism. Because they attend to what contemporary physics tells us, the overwhelming majority of contemporary compatibilists do not believe that determinism is true, but they do believe that even if it were true, people would be able to act freely. *Incompatibilism* is the thesis that free action is incompatible with the truth of determinism. In the incompatibilist group, most answers to the question what it is to act freely come from libertarians. *Libertarianism* is the conjunction of incompatibilism and the thesis that some people sometimes act freely. Some

incompatibilists argue that no one acts freely (Double 1991; Pereboom 2001; Strawson 1986). They argue that even the falsity of determinism creates no place for free action.

Compatibilist theories of free action emphasize a distinction between deterministic causation and compulsion (Frankfurt 1988; Smith 2003). If determinism is true, then my eating a banana for breakfast today and my working on this chapter today were deterministically caused; and so were a certain delusional person's spending the day trying to call the devil on her ham radio, a certain crack addict's using his favourite drug while in the grip of an irresistible urge to do so, a certain compulsive hand-washer's washing her hands scores of times today, and a certain person's giving his money to gunmen who convincingly threatened to kill him if he refused. But there is an apparent difference. I am sane and free from addiction, and I received no death threats today. The basic compatibilist idea is (roughly) that when mentally healthy people act intentionally in the absence of compulsion and coercion they act freely, and an action's being deterministically caused does not suffice for its being compelled or coerced.

Many compatibilists have been concerned to accommodate the idea that, for example, if I freely spent the day working, I could have done something else instead. They grant that, if determinism is true, then there is a sense in which people could never have done otherwise than they did: they could not have done otherwise in the sense that their doing otherwise is inconsistent with the combination of the past and the laws of nature. But, these compatibilists say, the fact that a person never could have done otherwise in that sense is irrelevant to free action. What is relevant is that people who act freely are exercising a rational capacity of such a kind that if their situation had been different in any one of a variety of significant ways, they would have responded to the difference with a different suitable action (Smith 2003). For example, although I spent the day working, I would have spent the day relaxing if someone had bet me $500 that I would not relax all day. This truth is consistent with determinism. (Notice that if someone had made this bet with me, the past would have been different from what it actually was.) And it reinforces the distinction between deterministic causation and compulsion. Offer a compulsive hand-washer $500 not to wash his hands all day and see what happens.

Like compatibilists, libertarians tend to maintain that when mentally healthy people act intentionally in the absence of compulsion and coercion they act freely, but they insist that determinism is incompatible with free action. Libertarians have the option of endorsing either stronger, non-historical requirements on free action or weaker, historical requirements (Mele 1995: 207–9). For example, they can claim that an agent freely *A*-ed at *t* only if, at *t*, he could have done otherwise than *A* then or claim instead that an agent who could not have done otherwise at *t* than *A* then may nevertheless freely *A* at *t*, provided that he earlier performed some relevant free action or actions at a time or times at which he could have done otherwise than perform those actions. I call any free *A*-ings that occur at times at which the past (up to those times) and the laws of nature are consistent with the agent's not *A*-ing then *basically free actions*. In principle, libertarians can hold that an agent's basically free actions that are suitably related to his subsequent *A*-ing confer freedom on his *A*-ing even though he could not have done otherwise than *A* then. It is open to libertarians to accept or reject the thesis that the only free actions are what I am calling basically free actions. These issues may safely be

skirted in this chapter; and to simplify exposition I announce that when libertarian freedom is at issue, by 'free' I mean 'basically free'.

Libertarian views of free action may be divided into three broad kinds, event-causal, agent-causal, and non-causal. According to typical event-causal libertarian views (e.g. Kane 1996), the proximate causes of free actions indeterministically cause them. This is a consequence of the typical event-causal libertarian ideas that free actions have proximate causes and that if an agent freely *A*-s at *t* in world *W*, he does not *A* at *t* in some other possible world with the same laws of nature and the same past up to *t*. Now, the proximate causes of actions, including actions that are decisions, are internal to agents. Even a driver's sudden decision to hit his brakes in an emergency situation is not proximately caused by events in the external world. Perception of whatever the source of the emergency happens to be—for example, a kangaroo darting into traffic—is causally involved. And how the driver decides to react to what he sees depends on, among other things, his driving skills and habits, whether or not he is aware of what is happening directly behind him, and his preferences. A driver who likes driving into kangaroos and is always looking for opportunities to do that would probably react very differently to the way a normal person would. In the light of the general point about the proximate causation of actions, typical event-causal libertarianism encompasses a commitment to what may be termed *agent-internal indeterminism.*

Agent-causal libertarianism features agent causation—causation of an effect by an agent or person, as opposed to causation of an effect by states or events of any kind, including a person's motivational and representational states (see e.g. Clarke 2003; O'Connor 2000). Think of causation as a relation between cause and effect. In ordinary event causation—for example, a lightning strike's causing a tree to crack—both cause and effect are events. These events are connected by the relation of causation. In agent causation, an agent is connected by the relation of causation to an effect.

Non-causal libertarianism, as Randolph Clarke (2003: 17) reports, 'imposes no positive causal requirement on free action'. Some views of this kind assert that only uncaused actions can be free (Goetz 2002), and others simply have no commitment to a causal account of free action (Ginet 2002; McCann 1998). If causalism about action is correct, there can be no uncaused actions—and, therefore, of course, no uncaused *free* actions. For an instructive critique of non-causal libertarian accounts of free action, see Clarke (2003: ch. 2).

Whereas the laws of nature that apply to deterministic causation are exceptionless, those that apply most directly to indeterministic causation are instead probabilistic.[11] Typically, events such as deciding to make a donation to Amnesty International—as distinct from the physical actions involved in actually making the donation—are counted as mental actions. Suppose that Ann's decision to make such a donation is indeterministically caused by, among other things, her thinking that she should donate money to that cause. Because the causation is indeterministic, she might not have decided to make a donation given exactly the same internal and external conditions. In this way, some libertarians seek to secure the possibility of doing otherwise that they require for free action.

This idea prompts a worry about luck that sets the stage for a discussion of the place of causation in event-causal and agent-causal libertarian views.[12] I illustrate the worry with a fable. Diana, a libertarian goddess in an indeterministic universe, wants to build rational, free human beings who are capable of being very efficient agents. She believes that proximal

decisions—typically, decisions to A straightaway—are causes of actions that execute them, and she sees no benefit in designing agents who have a chance of not even trying to A when they have decided to A straightaway and the intention to A that is formed in that act of deciding persists in the absence of any biological damage.[13] The indeterministic fabric of Diana's universe allows her to build a deterministic connection between proximal decisions and corresponding attempts, and she does. Now, because Diana is a relatively typical libertarian, she believes that free decisions cannot be deterministically caused—even by something that centrally involves a considered judgement that it would be best to A straightaway. She also believes that agents can make free decisions based on such judgements. So Diana designs her agents in such a way that, even though they have just made such a judgement, and even though the judgement persists in the absence of biological damage, they may decide contrary to it.

Given Diana's brand of libertarianism, she believes that whenever agents freely perform an action of deciding to A, they could have *freely* performed some alternative action. She worries that her design does not accommodate this. Her worry, more specifically, is that if the difference between the actual world, in which one of her agents judges it best to A straightaway and then, at t, decides accordingly, and any possible world with the same past up to t and the same laws of nature in which he makes an alternative decision while the judgement persists, is just a matter of luck, then he does not freely make that decision in that possible world, W. Diana suspects that the agent's making that alternative decision rather than deciding in accordance with his best judgement—that is, that difference between W and the actual world—is just a matter of bad luck, or, more precisely, of worse luck in W for the agent than in the actual world. After all, since the worlds do not diverge before the agent decides, there is no difference in them to account for the difference in decisions. This suspicion leads Diana to suspect that, in W, the agent should not be blamed for the decision he makes there.[14] And that he should not be blamed, she thinks, indicates that he did not freely make it. Diana has searched for grounds other than unfreedom for not blaming this agent and found none. She believes that if the agent had freely made the decision he made, he would have been morally responsible for it and blameworthy. Diana's worry about whether the agent's decision in W was freely made causes her to worry about whether his decision in the actual world was freely made, owing to the belief of hers reported in the opening sentence of this paragraph.

I am not privy to the details of Diana's design, but readers who require assistance in understanding her worry may consider the following story. As soon as any agent of hers judges it best to A, objective probabilities for the various decisions open to the agent are set, and the probability of a decision to A is very high. Larger probabilities get a correspondingly larger segment of a tiny indeterministic neural roulette wheel in the agent's head than do smaller probabilities. A tiny neural ball bounces along the wheel; its landing in a particular segment is the agent's making the corresponding decision. When the ball lands in the segment for a decision to A, its doing so is not *just* a matter of luck. After all, the design is such that the probability of that happening is very high. But the ball's landing there is *partly* a matter of luck. And the difference at issue at t between a world in which the ball lands there at t and a world with the same past and laws of nature in which it lands in a segment for another decision at t is just a matter of luck.

Diana can think of nothing that stops her worry from generalizing to all cases of deciding,

whether or not the agent makes a judgement about what it is best to do. In the actual world, Joe decides at t to A. In another world with the same laws of nature and the same past, he decides at t not to A. If there is nothing about Joe's powers, capacities, states of mind, moral character, and the like in either world that accounts for this difference, then the difference seems to be just a matter of luck. And given that neither world diverges from the other in any respect before t, there is no difference at all in Joe in these two worlds to account for the difference in his decisions. To be sure, something about Joe may explain why it is *possible* for him to decide to A in the actual world and decide not to A in another world with the same laws and past. That he is an indeterministic decision-maker may explain this. That is entirely consistent with the difference in his decisions being just a matter of luck. The worry just sketched is an instance of the problem of *present luck.*

All libertarians who hold that A's being a free action depends on its being the case that, at the time, the agent was able to do otherwise freely then should tell us what it could possibly be about an agent who freely A-ed at t in virtue of which it is true that, in another world with the same past and laws of nature, he freely does something else at t. Of course, they can *say* that the answer is 'free will'. But what they need to explain then is how free will, as they understand it, can be a feature of agents—or, more fully, how this can be so where 'free will', on their account of it, really does answer the question. To do this, of course, they must provide an account of free will—one that can be tested for adequacy in this connection (see sect. 4).

Incidentally, compatibilists face their own worry about luck. Incompatibilists want to know how agents can perform actions freely and be morally responsible for actions of theirs, if, relative to their own powers of control, it is just a matter of luck that long before their birth their universe was such as to ensure that they would perform those actions.[15] How, they want to know, is *remote deterministic luck* compatible with free action and moral responsibility.

4. LIBERTARIAN RESPONSES TO THE PROBLEM OF PRESENT LUCK

Robert Kane (1999) offers a detailed event-causal libertarian response to the worry about luck. He finds special importance in scenarios in which we struggle with ourselves about what to do (1996; 1999). In some cases of this kind, he says, we simultaneously try to make each of two competing choices or decisions (1999).[16] Since the agent is trying to make each, she is morally responsible for whichever of the two decisions she makes and makes it freely, Kane claims, provided that 'she endorse[s] the outcome as something she was trying and wanting to do all along' (ibid. 231–40, 233). Someone who takes this position can consistently hold that even if the agent's deciding to A, as she in fact did, rather than deciding to B, as she did at the same time in another world with the same past and laws of nature—that is, that difference, D, between the two worlds—is just a matter of luck, the agent decides freely and is morally responsible for her decision. At least, that is so if, as it seems, this agent's satisfying Kane's alleged sufficient condition for free and morally responsible decision is consistent with D's being just a matter of luck.[17] What matters, in Kane's view, is that the agent tries to make each decision (in both worlds) and endorses the outcome in the way just mentioned. If Kane is right, he has provided a successful solution to the worry about luck that I presented—at least in scenarios of a certain kind.

Part of the inspiration for Kane's position is the point that 'indeterminism [sometimes] functions as an obstacle to success without precluding responsibility' and freedom (ibid. 227). In one of his illustrations, 'an assassin who is trying to kill the prime minister … might miss because' his indeterministic motor control system leaves open the possibility that he will fire a wild shot. Suppose the assassin succeeds. Then, Kane says, he 'was responsible' for the killing 'because he intentionally and voluntarily succeeded in doing what he was *trying* to do—kill the prime minister'. It may be claimed, similarly, that the indeterminism in the scenario does not preclude the killing's being a free action. If these claims are true, they are true even if the difference between the actual world at the time of the firing and any wild-shot world that does not diverge from the actual world before that time is just a matter of luck.

As Clarke observes, Kane's point does not get him far, for the presumption of those who judge that the assassin freely killed the Prime Minister is that he *freely tried* to kill him (2002: 372–3): if we are told that perhaps the assassination attempt was not free, all bets are off. Kane does not claim that in cases of dual efforts to choose, the choices made are products of freely made efforts. Nor has he put himself in a position to claim this, for he has not offered an account of what it is for an effort to choose to *A* to be freely made. Thus, there is a salient disanalogy between cases like that of Kane's assassin and Kane's dual trying cases: there is no presumption that the dual efforts to choose are freely made. And if the agent's efforts to choose in a dual trying scenario—unlike the assassin's effort to kill the Prime Minister—are not freely made, it is hard to see why the choice in which such an effort culminates should be deemed free.

Readers who need help in appreciating this last point should imagine that a manipulator compels an agent, Ann, simultaneously to try to choose to *A* and to try to choose to *B*, where *A* and *B* are competing courses of action that, in the absence of manipulation, Ann would abhor performing. Imagine also that the manipulator does not allow Ann to try to choose anything else at the time and that the manipulation is such that Ann will endorse either relevant 'outcome as something she was trying and wanting to do all along'. The tryings are internally indeterministic, but Ann does not freely try to make the choices she tries to make. Apparently, whatever she chooses, she does not freely choose it—especially if the sort of freedom at issue is the sort most closely associated with moral responsibility. To be sure, in this scenario the unfreedom of the efforts is tied to serious monkey business. But take the monkey business away: if the efforts to choose still are not freely made, why should a corresponding choice count as free? The combination of trying and endorsement that Kane describes does not suffice for freely making the decision one makes: that combination is present in Ann's case. It may be claimed that this combination would turn the trick in the absence of monkey business. Once an argument for that claim is advanced, one can try to assess it. An argument by analogy from such cases as the assassin's will not fly as long as the disanalogy I mentioned is in place. One way to eliminate the disanalogy is to produce an acceptable account of the freedom of an effort to choose to *A* and show that efforts at work in some dual trying scenarios satisfy the account. Of course, one would also have to deal with the strangeness of the suggestion that rational free agents try to choose to *A* while also trying to choose another course of action that they know is incompatible with *A*-ing (on this problem, see Clarke 2002: 372; 2003: 88).

Timothy O'Connor regards the apparent failure of event-causal libertarian views as motivation for libertarians to embrace agent causation. 'Active power' is O'Connor's name

for the power exercised in agent causation. This power, he writes, 'is the power to freely choose one's course of action for reasons' (2000: 95, italics removed). Now, what is it about active power in virtue of which it is the power to choose freely for reasons? O'Connor says that 'exerting active power is intrinsically a direct exercise of control over one's own behaviour' (ibid. 61). In this, exerting active power is supposed to differ from what happens when allegedly free choices are made according to libertarians (e.g. Kane) who hold that such choices are indeterministically caused by internal states and events. O'Connor contends that, on these event-causal libertarian views, agents do not directly control the outcome. He writes (*O*): 'There are objective probabilities corresponding to each of the [possible choices], but within those fixed parameters, which choice occurs on a given occasion seems, as far as the agent's direct control goes, a matter of chance' (ibid. p. xiii, cf. 29). This resembles my suggestion that the difference at *t* between the actual world, in which Joe decides then to *A*, and any world with the same past and laws in which he instead decides then not to *A* is just a matter of luck. Of course, I made no exception for agents who exercise 'active power' or 'direct control'.

What is the connection between my suggestion and *O*? Consider a scenario O'Connor (ibid. 74) describes. Tim deliberates about whether to keep working or to take a break and decides to continue working. Obviously, it is not just a matter of chance that he decides to do that, and O'Connor does not claim that it is. After all, Tim had significant reasons and motivation to continue working, and he chose for those reasons. Also, in virtue of the fact just reported about Tim's motivation and, for example, the fact that he had no motivation at all for smashing his computer monitor with his coffee mug, it was much more likely that he would decide to continue working than that he would decide to smash his monitor with his mug. What O'Connor claims in *O* to be a matter of chance, 'as far as [Tim's] direct control goes' given event-causal libertarian views of the sort at issue, seems to be the following cross-world difference: Tim's choosing at *t* to continue working rather than choosing at *t* to do something else, as he does in some possible worlds with the same past and laws of nature.

O'Connor's own agent-causal view is supposed to avoid this consequence. Assume that Tim chose freely in the scenario under consideration. Then, on O'Connor's (ibid. 74) view, Tim 'had the power to choose to continue working or to choose to stop, where this is a power to cause either of these mental occurrences. That capacity was exercised at *t* in a particular way (in choosing to continue working), allowing us to say truthfully that Tim at time *t* causally determined his own choice to continue working.' Suppose that the position reported in the preceding two sentences is true. Why should we suppose that the following cross-world difference is not a matter of chance or luck: that Tim exercised the capacity at issue at *t* in choosing to continue working rather than in choosing to do something else, as he does in some possible worlds with the same past and laws of nature? Grant that Tim 'causally determined his own choice to continue working'. Why aren't the differences in his causal determinings at *t* across worlds with the same past and laws of nature a matter of chance or luck? Tim was able to causally determine each of several choices, whereas a counterpart who fits the event-causal libertarian's picture was able to make—but not to causally determine—each of several choices. If it is a matter of chance that the latter agent chooses to keep working rather than choosing to do something else, why is it not a matter of chance that the former agent causally determines the choice he causally determines rather than causally determining a choice to do

something else?

Perhaps O'Connor is thinking that the conceptual relation between control and chance is such that the fact that Tim exercised direct control over which choice he makes answers each of these questions. Should his readers find this thought persuasive? I do not see why. Even if the fact that Tim exercised direct control in choosing to continue working is incompatible with its being just a matter of luck that he chose to continue working, this does not show that a relevant cross-world difference between his exercising direct control 'in [this] particular way' (ibid.) and his exercising it in choosing to do something else is not just a matter of luck. The reader should bear two points in mind. First, as I explained, it is not just a matter of luck that Tim chose to keep working even on event-causal libertarian views. Second, O'Connor does not place cross-world differences in agents' doings out of bounds in the context of free will: in fact, such differences are *featured* in his objection from chance to event-causal libertarians.

O'Connor can say that even if, in scenarios featuring agent causation, cross-world differences of the kind I mentioned *are* matters of chance, that does not stand in the way of freedom and moral responsibility. In assessing the event-causal libertarian view advanced in Kane 1996, he contends that, owing to the specific kind of chanciness involved, 'the kind of control that is exercised is too weak to ground [the agent's] responsibility for which of the causal possibilities is realized' (ibid. 40). O'Connor can claim that the chanciness involved in whether an agent causally determines one choice rather than another is very different, and so different that the control that is exercised in Tim's causally determining his choice to continue working is sufficiently strong to ground his responsibility for that choice. But I find no argument for this thesis about chance and control in O'Connor (2000).

A critic may object that I have neglected O'Connor's claim that 'Active power is the power to *freely* choose one's course of action for reasons' (ibid. 95; emphasis altered). I may be asked to concentrate on the thought that O'Connor is arguing that 'the power to freely choose one's course of action'—the *F-power*, for short—is a possible power. It is true that he argues for this thesis, but he also contends that the *F-power* is the power of an agent directly to causally determine his choices—the *D-power*—a power that 'in suitable circumstances is freely exercised by the agent himself' (ibid. 72).[18] What I have not been able to ascertain is why it should be believed that (in whatever circumstances are deemed 'suitable') having the *D*-power suffices for having the *F*-power and exercising the *D*-power in choosing to *A* is sufficient for freely choosing to *A*. For I have been unable to ascertain why, for example, the crucial difference in causal determination at *t* between the actual world and any world with the same past and laws of nature in which, at *t*, Tim directly causally determines a choice to take a break is not just a matter of luck and unable, as well, to ascertain why, if this difference is just a matter of luck, Tim nevertheless freely chooses to continue working and is morally responsible for that choice. If O'Connor is legitimately to win converts to his view, he needs to lay this worry about luck to rest.

It may be suggested that even if agent-causal libertarianism is threatened by luck, it can handle luck better than event-causal libertarianism can. Derk Pereboom contends that although the latter view may, unlike compatibilism, 'provide leeway for decision and action', it does not provide 'enhanced control' (2001: 55; also see Clarke 2003: 220–1). On event-causal libertarian views, Pereboom argues, alleged free choices are 'partially random' events in the sense that 'factors beyond the agent's control [non-deterministically] contribute to their

production … [and] there is nothing that supplements the contribution of these factors to produce the events' (2001: 54, 48). What Pereboom calls 'enhanced control' is provided by supplementation of this contribution, and agent causationists can claim that *agents* supplement it in producing their free choices. A certain kind of agent causationism 'posits, as a primitive feature of agents, the causal power to choose without being determined by events beyond the agent's control, and without the choice being a truly random [i.e. uncaused] or partially random event … . In the best version of this position, free choices are identical to activations of this causal power' (2001: 55).[19] Voilà!

Assume that agent-causal libertarianism provides agents with a species of control that is not available in compatibilist and event-causal libertarian theories. Even then, it seems to be just a matter of luck that an agent exercised his agent-causal power at *t* in deciding to *A* rather than exercising it at *t* in any of the alternative ways he does in other possible worlds with the same past and laws of nature. In the light of this, one is entitled to worry that free choices are *not* identical to activations of the causal power that Pereboom describes—one that includes 'enhanced control'. The 'enhanced control' that he identifies leaves the worry intact.

Clarke (2003: 178) develops an 'integrated agent-causal account' according to which 'the exercise of freedom-level active control … consist[s] in causation of an action by mental events and by the agent'. Comparing an agent whose decision is indeterministically caused by events alone with a counterpart agent whose decision is 'brought about as characterized by an integrated agent-causal account' (ibid. 159), Clarke contends that the latter

exercised greater active control; he exercised a further power to causally influence which of the open alternatives would come about. In so doing, he was literally an originator of his decision, and neither the decision nor his initiating the decision was causally determined by events. This is why [he] is responsible for his decision, and why it was performed with sufficient active control to have been directly free. If this explanation is correct, then … the concept of agent causation is crucially relevant to the problem of free will. (ibid. 160)

Suppose, for the sake of argument, that there is an agent-causal power and that it is different from all event-causal powers that agents have. Suppose also that because it is a different power and has something to do with controlling one's actions, mixing it with agents' event-causal powers provides 'enhanced control'. Even with these suppositions in place, the difference identified in the preceding paragraph seems to be just a matter of luck. One can enhance a collection of powers that is not up to the task of securing a capacity for free and morally responsible action and get an enhanced collection that also is not up to that task. I may try to lift a weight using the power of my right arm alone and fail. I may try again, this time using in addition the further power of my left arm as well; and I may fail again, the combined powers not being up to the task. If the weight is a ton, the combined powers are not enough to give me even a ghost of a chance of lifting it. For all I have been able to ascertain, the combination of agent causation with indeterministic event causation is similarly inadequate.[20]

5. PARTING REMARKS

It is an interesting question whether a fully adequate, comprehensive theory of causation would enable philosophers to make significant progress on some of the issues I have discussed here: the dispute between causalists and non-causalists about action; how causalists should deal with deviant causal chains; whether causalism makes agents vanish; the dispute between compatibilists and incompatibilists about free will; and how libertarians should deal with the problem of present luck. I will not try to answer it here, but perhaps some of the background I have provided will assist some readers in applying their own position on causation to some of these issues.

FURTHER READING

Action: Davidson's seminal essays are to be found in his 1980. Mele (1997*b*) is a collection of influential articles on the philosophy of action. Mele (2003) is a monograph in the philosophy of action that includes critical discussions of many contemporary views on central topics in the field.

Free will: Recent important monographs include van Inwagen (1983) (an influential defence of incompatibilism); Fischer (1994) (an influential defence of a species of compatibilism); Kane (1996) (a defence of an event-causal libertarian view of free will); Pereboom (2001) (a defence of the view that we lack free will); and Clarke (2003) (a study of a variety of libertarian accounts of free will). Kane (2002) is a comprehensive collection of recent articles on free will.

REFERENCES

ALMEIDA, M., and BERNSTEIN, M. (2003). 'Lucky Libertarianism', *Philosophical Studies* 113: 93–119.

ANSCOMBE, G. E. M. (1963). *Intention*. 2nd edn. Ithaca, NY: Cornell University Press.

ARISTOTLE (1915). *Nicomachean Ethics*, ed. W. Ross. The Works of Aristotle 9. London: Oxford University Press.

AUDI, ROBERT (1997). 'Acting for Reasons', in Mele (1997*b*).

BEEBEE, H., and MELE, A. (2002). 'Humean Compatibilism', *Mind* 111: 201–23.

BISHOP, J. (1989). *Natural Agency*. Cambridge: Cambridge University Press.

BRAND, M. (1984). *Intending and Acting*. Cambridge, Mass.: MIT.

CLARKE, R. (2002). 'Libertarian Views: Critical Survey of Noncausal and Event-causal Accounts of Free Agency', in Kane (2002).

—— (2003). *Libertarian Accounts of Free Will*. Oxford: Oxford University Press.

DAVIDSON, D. (1963). 'Actions, Reasons, and Causes', *Journal of Philosophy* 60: 685–700.

—— (1980). *Essays on Actions and Events*. Oxford: Clarendon.

DOUBLE, R. (1991). *The Non-Reality of Free Will*. New York: Oxford University Press.

DRETSKE, F. (1988). *Explaining Behaviour: Reasons in a World of Causes*. Cambridge, Mass.: MIT.

FISCHER, J. M. (1994). *The Metaphysics of Free Will*. Oxford: Blackwell.

FRANKFURT, H. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

GINET, C. (1990). *On Action*. Cambridge: Cambridge University Press.

—— (2002). 'Reasons Explanations of Action: Causalist versus Noncausalist Accounts', in Kane (2002).

GOETZ, S. (2002). 'Alternative Frankfurt-Style Counterexamples to the Principle of Alternative Possibilities', *Pacific Philosophical Quarterly* 83: 131–47.

GOLDMAN, A. (1970). *A Theory of Human Action*. Englewood Cliffs: Prentice-Hall.

HAJI, I. (1999). 'Indeterminism and Frankfurt-type Examples', *Philosophical Explorations* 2: 42–58.

HARMAN, G. (1997). 'Practical Reasoning', in Mele (1997*b)*.

JACKSON, F. (2000). 'Psychological Explanation and Implicit Theory', *Philosophical Explorations* 3: 83–95.

—— and PETTIT, P. (1988). 'Functionalism and Broad Content', *Mind* 97: 381–400.

— — (1990). 'Program Explanation: A General Perspective', *Analysis* 50: 107–17.

KANE, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.

—— (1999). 'Responsibility, Luck and Chance: Reflections on Free Will and Indeterminism', *Journal of Philosophy* 96: 217–40.

—— (2000). 'The Dual Regress of Free Will and the Role of Alternative Possibilities', *Philosophical Perspectives* 14: 57–79.

—— (ed.) (2002). *The Oxford Handbook of Free Will*. Oxford: Oxford University Press.

McCANN, H. (1974). 'Volition and Basic Action', *Philosophical Review* 83: 451–73.

—— (1998). *The Works of Agency*. Ithaca, NY: Cornell University Press.

MELDEN, A. (1961). *Free Action*. London: Routledge & Kegan Paul.

MELE, A. (1992). *Springs of Action*. New York: Oxford University Press.

—— (1995). *Autonomous Agents*. New York: Oxford University Press.

—— (1997*a*). 'Agency and Mental Action', *Philosophical Perspectives* 11: 231–49.

—— (ed.) (1997*b*). *The Philosophy of Action*. Oxford: Oxford University Press.

—— (2003). *Motivation and Agency*. New York: Oxford University Press.

—— (2006). *Free Will and Luck*. New York: Oxford University Press.

—— and MOSER, P. (1997). 'Intentional Action', in Mele (1997*b*).

MILL, J. S. (1961). *A System of Logic*. 8th edn. repr. London: Longmans, Green.

NAGEL, T. (1986). *The View from Nowhere*. New York: Oxford University Press.

O'CONNOR, T. (2000). *Persons and Causes*. New York: Oxford University Press.

PEREBOOM, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

RYLE, G. (1949). *The Concept of Mind*. London: Hutchinson's University Library.

SEARLE, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.

SEHON, S. (1994). 'Teleology and the Nature of Mental States', *American Philosophical Quarterly* 31: 63–72.

SMITH, M. (2003). 'Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion', in S. Stroud and C. Tappolet (eds.), *Weakness of Will and Practical*

Irrationality. Oxford: Clarendon.

STRAWSON, G. (1986). *Freedom and Belief*. Oxford: Clarendon.

—— (1994). 'The Impossibility of Moral Responsibility', *Philosophical Studies* 75: 5–24.

THALBERG, I. (1977). *Perception, Emotion, and Action*. New Haven: Yale University Press.

—— (1984). 'Do Our Intentions Cause Our Intentional Actions?' *American Philosophical Quarterly* 21: 249–60.

THOMSON, J. (1977). *Acts and Other Events*. Ithaca, NY: Cornell University Press.

VAN INWAGEN, PETER (1983). *An Essay on Free Will*. Oxford: Clarendon.

VELLEMAN, J. D. (1989). *Practical Reflection*. Princeton: Princeton University Press.

—— (1992). 'What Happens When Someone Acts?' *Mind* 101: 461–81.

—— (2000). *The Possibility of Practical Reason*. Oxford: Oxford University Press.

WALLACE, R. J. (1999). 'Three Conceptions of Rational Agency', *Ethical Theory and Moral Practice* 2: 217–42.

WILSON, G. (1989). *The Intentionality of Human Action*. Stanford: Stanford University Press.

WITTGENSTEIN, L. (1953). *Philosophical Investigations*, trans. G.E.M. Anscombe. New York: Macmillan.

CHAPTER 26
CAUSATION AND ETHICS

CAROLINA SARTORIO

IN this chapter I examine potential applications of the concept of cause to some central ethical concepts, views, and problems. In particular, I discuss the role of causation in the family of views known as consequentialism, the distinction between killing and letting die, the doctrine of double effect, and the concept of moral responsibility.

My main aim is to examine the extent to which an appeal to the concept of cause contributes to elucidating moral notions or to increasing the plausibility of moral views. Something that makes this task interestingly complex is the fact that the notion of causation itself is controversial and difficult to pin down. As a result, in some cases the success of its use in moral theory hinges on how certain debates about causation are resolved. I will point to examples of this phenomenon as I proceed.

1. CAUSATION AND CONSEQUENTIALISM

Consequentialist theories are theories according to which the moral status of an act is exhausted by the moral status of its consequences. What are the consequences of an act? Numerous answers are possible. We could take them to be its causal consequences. Or we could take them to be its causal and logical consequences. Or we could take them to be the whole state of the world following the act. And so on. Different answers generate different versions of consequentialism.[1] Here I will discuss the causal answer in particular. I will dub it 'causal consequentialism'.

The content and implications of causal consequentialism depend on the nature of causation. The concept of cause appears to be restrictive in some ways and inclusive in others. It is restrictive in that consequences of an act that are too intimately related to it are not causal. For instance, an act of killing can be said to have the death of the victim as a consequence, but the relation between the two is presumably not causal.[2] For the fact that there was a killing *entails* the fact that there was a death, and an entailment is not a causal relation (consider: the fact that I entered the room at noon entails the fact that someone entered the room at noon, but it didn't cause it). Thus, according to causal consequentialism, the victim's death would not be a

consequence of the killing, and thus it would not contribute to the moral status of the killing. On the other hand, the concept of cause also appears to be quite liberal in that every act has many effects extending indefinitely into the future. Causal chains can be very complex and they can link events in unexpected ways.

These features of the concept of cause have interesting implications for causal consequentialism. I will examine some of them.

First, the restrictive side of causation has the following implication. Imagine that an assassin murders a child's parents. When he does, the child becomes an orphan. But the assassin's murder of the parents doesn't *cause* the child's orphanhood, for the fact that the child's parents were murdered *entails* the fact that the child is an orphan, and entailment is not causation. Thus causal consequentialism would imply that the child's orphanhood is not among the consequences of the act of murdering the parents. This seems wrong because we want to blame the assassin for murdering the parents and making the child an orphan as a result. On the basis of a similar example, David Sosa has argued that a consequentialist should embrace a broader concept of consequence (Sosa 1993).

Now, the causal consequentialist could reply that, although the assassin's murder of the parents didn't cause the child's orphanhood, some other act by the assassin did, for example, his pulling the trigger (note that the fact that he pulled the trigger doesn't entail the fact that the child became an orphan; after all, he could have missed).[3] Thus we can still blame the assassin for the child's orphanhood on these grounds. However, an important part of the objection survives. Causal consequentialism would still not entail that the child's orphanhood accounts for the moral status of the act of murdering the parents, for *that* act didn't cause the child's orphanhood. And it is plausible to suggest that the immorality of that act is explained, at least in part, by the fact that it has the child's orphanhood as a consequence. (Or, at least, it makes sense to suggest that a consequentialist would want to say this.) In addition, there might be cases where, although an agent is to blame for an outcome, *no* act by the agent can be said to be a cause of the outcome. I will discuss such cases in sect. 4.

We have examined reasons to believe that the concept of cause might be too narrow for consequentialism to draw on it. Now let us examine if there is also a sense in which it might be too broad.

A common objection (or family of objections) to consequentialism is based on the observation that any act has too many consequences, more than could plausibly matter to the moral status of the act. Call this objection 'the irrelevant consequences objection'. One version of this objection points to the fact that many consequences of acts are unforeseen (if you look far enough down the chain of events), and thus consequentialism doesn't provide a criterion that one can easily appeal to in deciding what to do (Mill (1863) anticipates this kind of objection to his view). Imagine, for instance, that an apparently morally good act indirectly influenced the time at which Hitler's parents conceived a child. It is initially implausible to suggest that the fact that the act led eventually to genocidal consequences makes the act wrong. There are two standard consequentialist responses to this version of the objection. First, the consequentialist could reply that we shouldn't expect the theory to play the role of a decision procedure; after all, consequentialism is a principle about the morality of acts, not about how we can come to know what the right moral act is. Therefore the act might be immoral, though we couldn't have known it at the time (see e.g. Bales 1971). Second, a

consequentialist might want to retreat to a version of consequentialism according to which the moral status of an act is determined, not by the whole set of consequences of the act, but only by its 'expected' consequences, that is, the consequences that the agent could reasonably expect to occur (see e.g. Jackson 1991).

Still, there is a version of the objection that survives both these moves. Imagine that I can choose between pulling the trigger of a gun and not doing so. If I do, a person will die; if I don't, then a hit man who is waiting in the background will shoot and the victim will still die. (This type of case is called a 'pre-emption' case in the causation literature.) Intuitively, under these circumstances, I ought not to shoot. But consequentialism seems to entail that I should be indifferent, given that the consequences of my shooting and my not shooting are the same. In this case, it is not the lack of foreseeability, or the remoteness, of the consequences of not shooting that makes them intuitively irrelevant, but it is the idea that they are not really 'due to the agent'. Thus the objection cannot be addressed by, for example, retreating to an expected consequences version of consequentialism.

How does causal consequentialism fare with respect to this version of the irrelevant consequences objection? At first sight, it might seem that it is not particularly well placed to answer it. For, as I have pointed out, according to causal consequentialism, *any* item in the causal chain flowing from an act would seem to be a consequence of it. However, there are at least two ways in which a causal consequentialist could try to resist this implication. I will examine the prospects of each of these attempts.

First, the causal consequentialist might want to rely on the notion of 'proximate causation' discussed in the philosophy of law. (Sinnott-Armstrong points to this possibility in his 2003, but says that it hasn't been developed as a version of consequentialism.) On this view, an agent is a cause of an outcome only if the outcome 'flows from his agency': the interventions of other agents or of unexpected natural phenomena can break up causal chains and rid an agent of causal responsibility for an outcome. (See Ch. 37 below.) This view entails that certain unexpected outcomes and other remote consequences of an act are not its genuine consequences. Also, it entails that the victim's death is not a consequence of my failure to shoot in the hit-man case, given that it occurs via the intervention of another moral agent. Thus this view escapes the irrelevant consequences objection, even on its more recalcitrant version. But the price paid is too high. Given the role that such factors as expectations, norms, agency, and so on play in this view, it is hard to square it with the idea that causation is an objective, natural, and 'out in the world' relation. Thus I suggest that we set this view aside and look for a more promising alternative.[4]

Fortunately for the causal consequentialist, there is a better approach: to reject the transitivity of causation. Some philosophers have done so (see e.g. Hitchcock 2001; Yablo 2002) but, as far as I know, no one has explicitly drawn on it to build a more plausible version of consequentialism. Rejecting the transitivity of causation could help the causal consequentialist in the following way: if causation isn't transitive, then the existence of long causal chains linking events doesn't imply that the events are causally related. On these grounds, a causal consequentialist could claim that not every event in the chain flowing from a given act contributes to the moral status of the act.

To be clear: rejecting the transitivity of causation doesn't amount to claiming that, when *c* causes *d* and *d* causes *e*, *c never* causes *e*, but only that it doesn't *sometimes* (and that whether

or not it does depends on the specific features of each case). However, rejecting transitivity can help the causal consequentialist answer the objection posed by the hit-man case in particular. For the hit-man case seems to be a scenario where transitivity fails, if it fails at all. Let me explain.

Someone who rejected the transitivity of causation could say the following about the hit-man case. If I refrain from shooting, my failing to shoot causes the hit man to shoot, and the hit man shooting, in turn, causes the victim to die; however, my failing to shoot doesn't cause the victim to die. For this is a case where transitivity fails: intuitively, I am only a cause of the person's death if I shoot, not if I don't. This is an independently plausible claim. It is also consistent with an objective, natural, and 'out in the world' conception of causation. In particular, it is not the existence of a second *agent* in the causal chain that makes us want to say that my failing to shoot isn't a cause. To see this, note that we can construct cases with the same causal structure but deprived of moral agents. Here is an example: yesterday's rain caused the existence of a puddle today; however, if it hadn't rained, the ensuing dryness would have caused a bowl with water to crack, and the water would have filled the hole in a similar way. Intuitively, the rain caused the puddle, but the absence of rain wouldn't have caused the puddle—although it would have caused the crack in the bowl, and the crack in the bowl would have caused the puddle. The hit man case shares this structure. So it is plausible to claim that, if I don't shoot the victim, the hit man causes his death but I don't.[5]

I conclude that a causal consequentialist would benefit from a conception of causation that rejected transitivity. The hit-man objection rests on the intuition that, although a killing would have occurred regardless of what I do, if I decide not to pull the trigger, I am not the one who 'does it' and thus I am not responsible for the death. As we have seen, by rejecting the transitivity of causation the causal consequentialist can capture this intuition without stepping outside the consequentialist framework. Note, however, that the ultimate success of such a project depends on whether causation does turn out to be a non-transitive relation, that is, it depends on how a certain debate in the metaphysics of causation is resolved.

In sum, causal consequentialism appears to have some advantages and some disadvantages. A potential disadvantage is that it ignores possible non-causal consequences of acts, such as logical consequences, that might be relevant to the morality of acts. On the other hand, a potential advantage is that, armed with a non-transitive concept of cause, causal consequentialism can rebut some serious objections to consequentialism in an appealing way. Ideally, consequentialism would benefit from a concept of consequence that combined the benefits of the causal concept with those of a broader notion.

2. CAUSATION AND KILLING/LETTING DIE

Some philosophers believe that killing is (other things being equal) morally worse than letting die (see e.g. Foot [1984] 1994 and Quinn [1989] 1994). Call this thesis 'KLD'. If KLD were true, it could be used to explain some ordinary moral judgements. For instance, it could explain why some people think that failing to give to charity is not wrong (or at least it is less bad than, say, sending poisoned food to third-world countries), on the grounds that it is merely letting people die. Also, it could account for a distinction drawn in contemporary medical practice between 'active' forms of euthanasia (such as injecting a terminal patient with a

lethal dose of a drug), which are generally regarded as impermissible, and 'passive' forms (such as withholding a certain medical treatment), which are generally regarded as permissible, on the grounds that active euthanasia is killing and passive euthanasia is letting die.

If there were a moral difference between killing and letting die, what could account for it? Here I will look at attempts to account for it in causal terms.

The simplest way to build a causal account of KLD is to start by grounding the killing/letting die distinction on the action/omission distinction, and to claim that the action/omission distinction coincides with the distinction between causes and non-causes. Roughly, the proposal would be that a killing is the causing of a death because a killing involves an action and actions are causes, whereas a letting die is not the causing of a death because a letting die involves an omission and omissions are not causes (see e.g. Callahan 1989).

There are several problems with this type of approach. The first problem, which I'll have to bypass here, is the identification of the killing/letting die distinction with the action/omission distinction. As it has been pointed out, it seems possible to kill by omitting to do something (as when a kidnapper kills his victim by starving him, or by failing to give him the insulin he needs) and, conversely, it seems possible to let die by doing something (as when a respirator is turned off to let a patient die a peaceful death). (See e.g. Quinn [1989] 1994.)

But, setting this objection aside, there are also problems with the claim that actions can be causes but omissions cannot. Whether omissions can be causes is a highly debated issue in the literature on causation (see Ch. 19 above). Again, then, the success of this proposal hinges on how a certain debate in the metaphysics of causation is resolved. More importantly, however, the following dilemma arises for someone who wants to take this line. Common sense dictates that at least some omissions are causes. Thus, on the one hand, if common sense is right and omissions can be causes, the appeal to the action/omission distinction doesn't help account for KLD. On the other hand, suppose that common sense is wrong and omissions cannot be causes. Then, I take it, the reason the proposal might seem promising to anyone is that, intuitively, there is a clear moral distinction between causing deaths and not causing deaths. That is to say, the thought would have to be that, intuitively, causing harm is bad (at least under certain circumstances) and not causing harm isn't, so killing someone can be bad because it is the causing of a death but letting someone die isn't bad because it isn't the causing of a death. However, this is too strong: surely, one can merely let someone die and still act very badly. For instance, if I see a seriously injured person on my way home and I don't stop to save him to prevent my take-out pizza from getting cold, I don't kill the person (I merely let him die) but I still act wrongly. Consequently, it seems that, if one embraced the view that omissions cannot be causes, then one would have to acknowledge that not causing harm can be bad. As a result, the project of grounding KLD would still fail, for then the connection between not causing harm and acting permissibly (or less wrongly) would be severed. (For an argument along the lines of this second horn, see Howard-Snyder 2002.)

In sum, the dilemma that a causal account of KLD would have to face is:

    (Premiss)   Either omissions can be causes or they cannot.

(Horn 1)	If omissions can be causes, then there is no causal difference between actions and omissions, and thus we cannot account for the killing/letting die distinction in causal terms.
(Horn 2)	If omissions cannot be causes, then we cannot ground the distinction between killing and letting die in the claim that causing harm is bad but not causing harm isn't, and thus we cannot account for the killing/letting die distinction in causal terms.
(Conclusion)	We cannot account for the killing/letting die distinction in causal terms.

Now, one could try to resist the dilemma by rejecting horn 1. There are two main ways of doing this. First, one could claim that some omissions can be causes, but not all of them can (in particular, not those involved in letting-die cases). Second, one could argue that, even if all omissions can be causes, omissions have importantly different causal powers from those of actions (and those different causal powers generate the moral difference between killing and letting die). Note that these views would still be causal accounts of KLD. In what follows I briefly examine the prospects of each view.

Let us examine, first, the view that some omissions have causal powers but others don't. Presumably, the proposal would be that, for example, when a kidnapper starves his victim to death, his failing to feed him causes the victim's death; by contrast, when a doctor fails to continue a medical treatment on a terminal patient, his discontinuing the treatment does not cause the patient's death (instead, the patient's ailment does). On this view, such causal difference explains why the doctor merely lets his patient die, and thus acts permissibly, whereas the kidnapper kills his victim, and thus acts impermissibly.

However, there is a serious objection to this view. What the view does, at bottom, is to attribute causal powers only to those omissions that are *salient* in the circumstances. The most salient threat to the patient is his ailment (not the doctor); the most salient threat to the kidnapped victim is the kidnapper. Thus, whereas the doctor's omission doesn't strike us as a cause, the kidnapper's omission does. However, salience is only a pragmatic factor. Pragmatic factors cannot make a metaphysical difference unless they are accompanied by other metaphysically relevant factors. Now, one could try to argue that in the doctor case there is a pre-existing physical process (the disease) that, if left alone, will lead to the death, but in the kidnapper case there isn't such a process. If this were true, then there would be a metaphysical difference between the cases that could account for a causal difference. However, it is false: in the kidnapper case too, there are some biological processes that will lead to the victim's death if he doesn't get the nutrients he needs to stay alive.

The difficulty encountered by this type of account of KLD is reminiscent of the 'Queen of

England problem' in the literature on causation by omission (see e.g. Beebee 2004). The Queen of England problem is this: imagine that we want to claim that the gardener's failure to water my plant was a cause of the plant's death. Then it seems that we'll have to acknowledge that everyone else's failure—including the Queen of England's—was a cause too. After all, there is no relevant metaphysical difference between the gardener's failure and the Queen of England's failure: the only ways in which they differ concern duties, expectations, etc. As a result, it seems that we should believe that the Queen of England causally contributed to my plant's death. The standard attitude in the literature (by those who accept the possibility of causation by omission) is to bite the bullet and accept this result (see e.g. Lewis 2004).

We have seen that the view that assigns different causal powers to different omissions faces serious difficulties. Now let us examine the view that actions (in general) and omissions (in general) have importantly different causal powers, the second suggested way of escaping the dilemma presented above. Roughly, this view claims that, although actions and omissions can both be causes, they cause things in importantly different ways, and this difference in causal powers is significant enough to generate a moral difference between killing and letting die.

What could be the difference in causal powers between actions and omissions? Something that comes to mind is the distinction that some philosophers have drawn between 'enabling' conditions (or 'enablers') and 'triggering' conditions (or 'triggers') (see e.g. Lombard 1990). Roughly, an enabler is something that 'facilitates' the occurrence of an effect, or merely makes it possible, without setting off the causal chain leading to it. Enablers are sometimes regarded as 'background conditions': facts or states of affairs that need to be in place for an outcome to happen, but that require a triggering event for the causal chain to unfold. On some views, enablers are not genuine causes (Lombard 1990; Thomson 2003). But one could also argue that enablers *are* causes, although their contribution is different from that of triggers. For instance, one might want to distinguish the causal contribution of an enabler, such as the oxygen present in the environment, without which a match wouldn't have lit, from that of a trigger such as the striking of the match. Similarly, one might want to distinguish the contribution of the mere existence of a bridge to a person's act of crossing it from that of the reasons (beliefs and desires) for which he crossed it (I am adapting one of Thomson's examples in her 2003).

On the basis of the distinction between enablers and triggers, the next step would be to argue that omissions are always enablers, never triggers, and that this causal difference gives rise to a moral difference. For instance, it could be argued, the doctor's failure to continue the treatment is only an enabler in that it only 'facilitates' the occurrence of the patient's death, with the ailment being the trigger. Then the suggestion would be that, just like we normally regard the act of pulling a trigger as worse than the act of leaving a gun on a table unattended, similarly, we should regard an act of killing as worse than an act of letting die.

This view has two serious problems. First, at least sometimes, omissions—and absences in general—play the role of triggers. For instance, the kidnapper's failure to feed his victim 'triggers' the victim's death, and the absence of rain 'triggers' the drought. Thus the identification of omissions with enablers fails. Second, almost anything can be seen as a mere enabler. Take the striking of the match. Other things had to happen, besides the striking of the match, for it to light. So the striking of the match too, just like the presence of oxygen in the environment, is something that merely 'facilitates' the occurrence of the effect, at least in the

sense that it is not independently sufficient for the occurrence of the effect. Similarly, in the passive euthanasia case, it is confused to suggest that the ailment is independently sufficient for the death but the doctor's omission isn't: they need each other (as well as other conditions) to bring about the death. Again, it seems that the reason why we are tempted to draw a distinction between them is that one of them is more salient than the other, but salience isn't enough to ground a metaphysical difference. (Penelope Mackie raises a similar objection to Lombard's view of enablers in her 1992.)

I conclude that there are serious problems with the project of accounting for KLD in causal terms. The project faces a dilemma from which it is difficult to escape. Hence, if there is a moral difference between killing and letting die, it probably doesn't have a causal basis.

3. Causation and Double-Effect

On a classical formulation, the 'doctrine of double effect' (DDE) says that an act that issues in a bad result is permissible if four conditions are met: first, the act is not bad in itself; second, the act also issues in a proportionally good result; third, the bad result is not intended but merely foreseen; fourth, the bad result is not a means to the good result (Mangan 1949).

It is natural to understand the fourth condition—hereafter *the means condition*—causally: on this interpretation, the bad result is not a means to the good result when the bad result doesn't *cause* the good result. In turn, it is natural to interpret the third condition—hereafter *the intentionality condition*—as the subjective or psychological counterpart of the means condition. That is to say, if the means condition reads: 'The bad result is not a causal means to the good result,' the intentionality condition reads: '*The agent* doesn't view the bad result as a causal means to the good result.' This is the version of DDE that I will be concerned with here. Some versions of DDE don't explicitly appeal to the concept of cause; instead, they make primitive use of the means–end distinction or they analyse this distinction in other terms. I will argue that a *causal* version of DDE has an important advantage over other versions in that it can provide an appealing answer to a common objection to DDE.

Let us start by looking at some examples. DDE is generally used to account for the moral difference between such cases as these:

> *Terror Bomber*: A pilot bombs civilians in order to lower the enemy's morale and thus end the war. The end of the war saves many lives.

> *Strategic Bomber*: A pilot bombs a weapons factory in order to end the war. The bombing issues in the end of the war, but also in some civilian deaths. (Bratman 1987)

Intuitively, the terror bomber acts impermissibly but the strategic bomber acts permissibly. Proponents of DDE claim that the difference consists in that, whereas the terror bomber intends the civilian deaths, or uses those deaths as means to his end, the strategic bomber doesn't: the civilian deaths are merely a foreseen side-effect of the strategic bomber's act and are not means to his end.

Also, DDE is often used to explain the difference between these two cases:

> *Transplant*: A surgeon kills a healthy patient in order to use his organs to save five people. The one dies but the five live.

> *Trolley*: A runaway trolley is hurtling down the tracks where five people are trapped. A bystander flips a switch that redirects the train towards a side track, where only one person is trapped. The one dies but the five live.

Intuitively, the surgeon in Transplant acts impermissibly but the bystander in Trolley acts permissibly. The problem of accounting for this difference has been called 'the trolley problem' (cf. Foot [1967] 1994 and Thomson [1976] 1986). Proponents of DDE suggest this solution to the problem: in Transplant, the surgeon intends the death of the one (without which he couldn't save the five) and uses his death as a means to saving the five, but in Trolley the bystander doesn't intend the death of the one and his death is not a means to the saving of the five (instead, it is a mere side-effect).

Now, a common objection to DDE is that the intended/foreseen distinction, as well as the distinction between one's means and what results from one's means, is fundamentally unclear (see e.g. Davis 1984). To illustrate the problem, consider Terror Bomber again. The advocate of DDE wants to claim that the terror bomber acts impermissibly because, unlike the strategic bomber, he uses the civilian deaths as means to his end and the deaths are an intended consequence of his dropping the bombs. However, we are assuming that the terror bomber only bombs the civilians because he wants the war to come to an end: he doesn't do it just to kill some civilians. If there were a way for him to avoid their deaths and at the same time achieve his desired end (say, by misleading the enemy into thinking that the civilians are dead, thus still demoralizing them), then he would do it. This suggests that he doesn't really intend the civilians to die, but only to *seem* dead. Also, one could argue that in Terror Bomber the civilian deaths are not really a means to the war's coming to an end. The war would still have ended if the civilians had only seemed dead; hence, the means to the end of the war was only their seeming dead. Thus, the objection goes, DDE fails to explain the moral difference between Terror Bomber and Strategic Bomber.

The same objection could be raised for Transplant and Trolley. One could argue that, in Transplant, if the surgeon could somehow get the one patient's organs without killing him, he would do it. Hence, the surgeon doesn't really intend the one's death, but only the removal of his organs. Also, the means to the five's survival was not the one's death but the removal of the organs. Thus, the objection goes, DDE fails to explain the moral difference between Transplant and Trolley.

I think that the advocate of a *causal* version of DDE can answer this objection in a way that isn't available to other versions of the view. He could argue as follows. The truth of the counterfactual:

> Had the civilians seemed dead without being dead, the war would still have ended.

doesn't entail that the civilian deaths didn't cause the end of the war. In the actual case, the deaths did cause the end of the war, via their causing it to be the case that the civilians seemed dead. Similarly, the truth of the counterfactual:

> Had the surgeon removed the organs from the one without killing him, the five would still have survived

doesn't entail that the one's death didn't cause the five's survival. In the actual case, the one's death did cause the five's survival, via its facilitating the removal of the organs from the one's body.[6]

In other words, the objection only works under this assumption:

> *Counterfactual conception of means*: If the counterfactual 'Had X occurred without Y, Z would still have occurred' is true, then Y is not a means to Z.

But, as it has been pointed out in the literature on causation, the concept of cause doesn't seem to meet this simple counterfactual requirement. Consider a case of pre-emption: Fast Shooter shoots at Victim and Victim dies as a result; Slow Shooter also shoots but it takes longer for his bullet to arrive, so he is pre-empted by Fast Shooter. In this case, the counterfactual 'Had Slow Shooter shot without Fast Shooter shooting, Victim's death would still have occurred' is true. However, Fast Shooter is still a cause of Victim's death. This suggests that, if the notion of means is understood causally, as opposed to, for example, merely counterfactually, the objection to DDE fails. (Again, this would depend, ultimately, on how the debate over the relation between causation and counterfactuals is resolved. But, at least initially, it seems that pre-emption cases undermine counterfactual theories of causation on their simplest versions. See Ch. 8 above.)

I have formulated the reply to the objection in terms of the causal condition. But, arguably, it can be extended to the intentionality condition as well, which (as I am understanding it) is simply the psychological counterpart of the causal condition. Thus, an advocate of a causal version of DDE could claim that, in Terror Bomber, the pilot *does* intend the civilian deaths. To be sure, his goal *could* still have been achieved by the civilians only seeming dead. But in the actual case it wasn't: in the actual case, the deaths caused the war to come to an end, and the terror bomber viewed those deaths as a causal means to his goal. Similarly, in Transplant, the surgeon *does* intend the one's death. Again, his goal *could* still have been achieved by the one's surviving the removal of his organs, but in the actual case it wasn't: in the actual case his goal was achieved by the one's death causing the saving of the five.

I have argued that a causal version of DDE has an important advantage over other versions of the doctrine. Namely, a causal interpretation of the notion of means, and of the related

concept of an intended consequence, provides the doctrine with a promising answer to a standard objection.

4. CAUSATION AND MORAL RESPONSIBILITY

In this section I examine the relationship between causation and moral responsibility. I will not touch on issues that are the focus of another chapter, such as the relation between responsibility and causal determinism or the concept of 'agent causation'. (See Ch. 25 above.)

Causation seems to be related to moral responsibility in this way: agents are normally regarded as responsible for (some of) their actions and omissions but also for (some) external events and states of affairs. Which events and states of affairs? A natural suggestion is that the *only* events and states of affairs that agents are responsible for are certain causal products of their actions and omissions. After all, the causal powers of an agent's actions and omissions seem to constitute the only links between the agent and the world in virtue of which the agent can impact the world. Call this 'the received view' about the relation between moral responsibility (for events and states of affairs in the world) and causation. According to the received view, again, an agent is morally responsible for something *only if* the agent caused it, that is, only if some action or omission of his caused it. Plausibly, other conditions are required for the agent to be responsible, for example, the agent could or should have foreseen that the outcome would result from his behaviour, but the causal condition is usually taken to be a necessary condition. (For an example of a theory of responsibility that contains the causal condition as a component, see Feinberg 1970.) Here I will focus on the causal condition only. I will critically examine possible reasons to resist it.

How could an agent be responsible for something without causing it? Someone might think that the Sosa-type example from sect. 1 shows a way. Thus, one might think that an assassin can be responsible for a child's orphanhood in virtue of having killed the child's parents even though his killing the parents didn't *cause* the child's orphanhood. However, as I have argued in sect. 1, there are other actions by the assassin that arguably did cause the child's orphanhood, for example, his pulling the trigger of his gun. As a result, the causal condition is unscathed by this example: the assassin caused the child's orphanhood, since *some* act of his caused it.

I will argue, however, that there are other cases that threaten to undermine the causal condition. In what follows I put forth two such cases: one involving omissions and one involving commissions.

Case 1 (omissions): imagine that two buttons have to be depressed at the same time to prevent a bad outcome from happening. Imagine that the agents in charge of the buttons independently fail to depress them because they want the harm to ensue. In that case, I submit, each agent is responsible for the outcome but neither agent caused it, in particular, neither of the individual omissions by the agents caused it. Briefly, the argument that the agents' omissions didn't cause the outcome goes thus: imagine that there had only been one agent: he was in charge of one button and an automated mechanism was in charge of the other button. In that case, if the mechanism failed, the agent's failure wouldn't be a cause (after all, there was nothing that the agent could have done to prevent the outcome, since the mechanism failed and

both buttons needed to be depressed to prevent it). But whether the other button was being controlled by an automated mechanism or by another agent cannot possibly matter to the causal powers of the agent. So, in the case with two agents, neither agent's failure is a cause of the outcome. But, clearly, each agent bears some responsibility for the outcome. Thus it is possible to be responsible for an outcome without causing it. (For a more extensive discussion, see Sartorio 2004.)

Case 2 (commissions): imagine that a person is tied to a train track and that a runaway train is approaching him. There is a switch and a side track diverging from the main line, but the tracks reconverge before the spot where the person is located. Thus the train can kill the person via either route, assuming that the relevant pieces of track are in working order. Now imagine that there are two agents, Flipper and Reconnecter. Flipper is by the switch and flips it when the train approaches it; as a result, the train runs on the side track for a while, then on the main track again, and ends up killing the person. Reconnecter is by the segment of the main track where the tracks come apart. As it turns out, that piece of track had been disconnected earlier that day. As a result, if the train had run on the disconnected piece of track, it would have derailed and it wouldn't have killed the person. Reconnecter reconnects that piece of track at around the same time that Flipper flips the switch.

I submit that this is another example of responsibility without causation. (For a more extensive discussion, see Sartorio (2006).) Note, first, that, had it not been for Flipper and Reconnecter, the person wouldn't have died: the train would have continued on the main track and it would have derailed while passing through the disconnected segment. This is just like in Case 1, where the ensuing harm depends on the combined behaviour of two moral agents. Thus, assuming that Flipper and Reconnecter were fully aware of what they were doing, it is likely that we want to blame them for the person's death. In particular, focus on Flipper's behaviour: it seems clear that we want to blame him, at least partly, for the death. However, I submit that his flipping the switch isn't a cause of the person's death. The argument for this parallels the earlier argument for Case 1. Imagine that what reconnected the segment of the main track had been an automated mechanism; in that case we wouldn't want to say that Flipper's redirection of the train was a cause of the person's death. After all, what he did was only to redirect a threat from one path to another, where, given the presence of the reconnecting mechanism, neither path was more threatening than the other.[7] But whether what reconnected the track was a mechanism or a moral agent cannot determine whether Flipper was a cause. So Flipper isn't a cause when what reconnects the track segment is Reconnecter. Thus Flipper is responsible for the death without causing it.

I have argued that there is reason to believe that moral responsibility and causation are not related in the way suggested by the received view. That is to say, being morally responsible for an outcome doesn't require causing it. How are they related, then? I will offer the sketch of a suggestion. In Case 1, even if the two agents do not cause the harm, they are responsible for something that causes the harm. What causes the harm? The fact that the two buttons were not simultaneously depressed (since, had they *both* been depressed, the harm wouldn't have occurred). The two agents were responsible for this fact, and thus, given that *this* fact caused the harm, they are responsible for the harm. Something similar is true of Case 2. Even though Flipper and Reconnecter don't cause the harm, they are responsible for something that causes it. What causes the harm? The fact that the train ran on a viable track (since, had it not run on

a viable track, it wouldn't have reached the person). Flipper and Reconnecter are responsible for this fact and, given that *this* fact caused the harm, they are responsible for the harm. Now, this proposal gives rise to questions that I will have to set aside here, such as the threat of a possible regress (for example, how are Flipper and Reconnecter responsible for the further fact that the train ran on a viable track? Did they cause *that* fact?). My aim here has been only to make plausible the idea that the received view about the relation between responsibility and causation needs revision, and to offer a rough suggestion as to how to revise it. (For discussion of the alternative view, see Sartorio 2004.)

FURTHER READING

On consequentialism: Classic discussions include Mill (1863), Bentham ([1789] 1996), and Moore (1903); a useful anthology is Scheffler (1988).

On killing and letting die: The distinction is defended in, for example, Dinello ([1971] 1994) and Kamm (1983). Some criticisms of the distinction are to be found in Bennett (1995) and Rachels ([1975] 1994).

On double-effect: Classic pieces include Aquinas (1988), Anscombe (1981), and Mangan (1949). A good anthology is Woodward (2001).

On moral responsibility: Classic papers include Strawson (1962), Frankfurt (1971), and Fischer and Ravizza (1998).

On the role of causation in ethics: Thomson (1990: ch. 5) argues that the notion of consequence presupposed by consequentialism is not causal and offers an alternative account. McGrath (2003) offers an account of the killing/letting die distinction in terms of causal structures. Bennett (1995) argues that the causal condition is not an essential component of DDE. *Pace* Bennett, McIntyre (2005) regards the causal condition as an essential element of DDE. Finally, Dowe (2001) argues that omissions can only be 'quasi'-causes, and that quasi-causation can make us as morally responsible for outcomes as genuine causation does.

REFERENCES

ANSCOMBE, G. E. M. (1981). 'Mr. Truman's Degree', in *Collected Philosophical Papers of G. E. M. Anscombe*. Oxford: Blackwell, iii. 62–71.

AQUINAS, T. (1988). *Summa Theologica II-II*, q. 64 a. 7, 'Of Killing', in W. P. Baumgarth and R. J. Regan (eds.), *On Law, Morality, and Politics*. Indianapolis: Hackett, 226–7.

BALES, R. E. (1971). 'Act-Utilitarianism: Account of Right-Making Characteristics or Decision Procedures?' *American Philosophical Quarterly* 8: 257–65.

BEEBEE, H. (2004). 'Causing and Nothingness', in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 291–308.

BENNETT, J. (1995). *The Act Itself*. New York: Oxford University Press.

BENTHAM, J. ([1789] 1996). *An Introduction to the Principles of Morals and Legislation*. Oxford: Clarendon Press.

BRATMAN, M. (1987). *Intentions, Plans, and Practical Reason*. Cambridge, Mass.: Harvard University Press.

CALLAHAN, D. (1989). 'Killing and Allowing to Die', *The Hastings Center Report* 19.

DAVIDSON, D. ([1971] 1980). 'Agency', originally in R. Binkley, R. Bronaugh, and A. Marras (eds.), *Agent, Action, and Reason*. Toronto: University of Toronto Press; repr. in *Essays on Actions and Events*. Oxford: Oxford University Press, 43–61.

DAVIS, N. (1984). 'The Doctrine of Double Effect: Problems of Interpretation', *Pacific Philosophical Quarterly* 65: 107–23.

DINELLO, D. ([1971] 1994). 'On Killing and Letting Die', *Analysis* 31; repr. in Steinbock and Norcross (1994: 192–6).

DOWE, P. (2001). 'A Counterfactual Theory of Prevention and "Causation" by Omission', *Australasian Journal of Philosophy* 79: 216–26.

FEINBERG, J. (1970). 'Sua Culpa', in *Doing and Deserving: Essays in the Theory of Responsibility*. Princeton: Princeton University Press, 187–221.

FELDMAN, F. (1997). *Utilitarianism, Hedonism, and Desert*. New York: Cambridge University Press.

FISCHER, J. M., and RAVIZZA, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. New York: Cambridge University Press.

FOOT, P. ([1967] 1994). 'The Problem of Abortion and the Doctrine of the Double Effect', originally in *Oxford Review* 5; repr. in Steinbock and Norcross (1994: 266–79).

—— ([1984] 1994). 'Killing and Letting Die', originally in J. L. Garfield and P. Hennessey (eds.), *Abortion: Moral and Legal Perspectives*. Amherst: University of Massachusetts Press; repr. in Steinbock and Norcross (1994: 280–9).

FRANKFURT, H. (1971). 'Freedom of the Will and the Concept of a Person', *Journal of Philosophy* 68: 5–20.

HITCHCOCK, C. (2001). 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98: 273–99.

HOWARD-SNYDER, F. (2002). 'Doing vs. Allowing Harm', *The Stanford Encyclopedia of Philosophy (Summer 2002 Edition)*, ed. Edward N. Zalta; http://plato.stanford.edu/archives/sum2002/entries/doing-allowing/, accessed 18 March 2009.

JACKSON, F. (1991). 'Decision-Theoretic Consequentialism and the Nearest and Dearest Objection', *Ethics* 101: 461–82.

KAMM, F. (1983). 'Killing and Letting Die: Methodological and Substantive Issues', *Pacific Philosophical Quarterly* 64: 297–312.

LEWIS, D. (1986). 'Events', *Philosophical Papers II*. Oxford: Oxford University Press, 241–69.

—— (2004). 'Causation as Influence', in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 75–106.

LOMBARD, L. (1990). 'Causes, Enablers, and the Counterfactual Analysis', *Philosophical Studies* 59: 195–211.

MCGRATH, S. (2003). 'Causation and the Making/Allowing Distinction', *Philosophical Studies* 114: 81–106.

MCINTYRE, A. (2005). 'Doctrine of Double Effect', *The Stanford Encyclopedia of Philosophy (Summer 2005 Edition)*, ed. Edward N. Zalta; http://plato.stanford.edu/archives/sum2005/entries/double-effect/, accessed 18 March 2009.

MACKIE, P. (1992). 'Causing, Delaying, and Hastening: Do Rains Cause Fires?' *Mind* 101: 483–500.

MANGAN, J. (1949). 'An Historical Analysis of the Principle of Double Effect', *Theological Studies* 10: 41–61.

MILL, J. S. (1863). *Utilitarianism.* London: Parker, Son & Bourn.

MOORE, G. E. (1903). *Principia Ethica.* Cambridge: Cambridge University Press.

QUINN, W. ([1989] 1994). 'Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing', originally in *Philosophical Review* 98; repr. in Steinbock and Norcross (1994: 355–82).

RACHELS, J. ([1975] 1994). 'Active and Passive Euthanasia', *The New England Journal of Medicine* 292; repr. in Steinbock and Norcross (1994: 112–19).

ROWE, W. (1989). 'Causing and Being Responsible for What is Inevitable', *American Philosophical Quarterly* 26: 153–9.

SARTORIO, C. (2004). 'How to be Responsible for Something without Causing It', *Philosophical Perspectives* 18: 315–36.

—— (2005). 'Causes as Difference-Makers', *Philosophical Studies* 123: 71–96.

—— (2006). 'Disjunctive Causes', *Journal of Philosophy* 103: 521–38.

SCHEFFLER, S. (ed.) (1988). *Consequentialism and its Critics.* Oxford: Oxford University Press.

SINNOTT-ARMSTRONG, W. (2003). 'Consequentialism', *The Stanford Encyclopedia of Philosophy (Summer* 2003 *Edition)*, ed. Edward N. Zalta http://plato.stanford.edu/archives/sum2003/entries/consequentialism/, accessed 18 March 2009.

SOSA, D. (1993). 'The Consequences of Consequentialism', *Mind* 102: 101–22.

STEINBOCK, B., and Norcross, A. (eds.) (1994). *Killing and Letting Die.* New York: Fordham University Press.

STRAWSON, P. F. (1962). 'Freedom and Resentment', *Proceedings of the British Academy* 48: 1–25.

THOMSON, J. ([1976] 1986). 'Killing, Letting Die, and the Trolley Problem', originally in *The Monist* 59; repr. in William Parent (ed.), *Rights, Restitution, and Risk: Essays in Moral Theory*. Cambridge, Mass.: Harvard University Press, 78–93.

—— (1990). *The Realm of Rights*. Cambridge, Mass.: Harvard University Press.

—— (2003). 'Causation: Omissions', *Philosophy and Phenomenological Research* 66: 81–103.

WOODWARD, P. A. (ed.) (2001). *The Doctrine of Double Effect: Philosophers Debate a Controversial Moral Principle*. Notre Dame, Ind.: University of Notre Dame Press.

YABLO, S. (2002). 'De Facto Dependence', *Journal of Philosophy* 99: 130–48.

CHAPTER 27
CAUSAL THEORIES OF KNOWLEDGE AND PERCEPTION

RAM NETA

PHILOSOPHERS have made many attempts to explain what it is for one thing to cause another thing; causation has thus often been the target explanandum of philosophical explanation.

But it is also true that causation has often been part of the philosophical explanans of other target explanantia. For instance, philosophers have often appealed to causation in order to explain what it is for someone to perceive a thing that exists independently of her perception or awareness of it—what we shall henceforth call an 'external thing'. And philosophers have often appealed to causation in order to explain what it is for someone to know about external things by means of perception. The phrase 'causal theory of perception' has been applied to theories of both of these two kinds. The phrase 'causal theory of knowledge' has been applied to theories of the second kind only.

This chapter will survey theories of both of these two kinds. We will first survey those 'causal theories of perception' that attempt to explain what it is for someone to perceive an external thing. Then we will survey the other 'causal theories of perception' (or alternatively, 'causal theories of knowledge') that attempt to explain what it is for someone to know about external things by means of perception. Within each of these two topics, we can locate all the various causal theories on a two-dimensional map: along one dimension are the various things that have been taken to do the causing, and along the other dimension are the various things that have been taken to be thus caused.

Why consider theories of these two kinds in one chapter? Is it simply because the phrase 'causal theory of perception' has, as a matter of (potentially confusing) historical accident, been applied to both kinds of theories? Of course, the question of what it is for someone to perceive a thing is distinct from the question of what it is for someone to know about external things by means of perception. But, though these questions are distinct, they are also very closely related. They are related because perceiving an external thing is, at least for creatures like us, either a way of knowing, or else a way of coming to know, about that thing by means of perception. If there is a causal component to perceiving an external thing, then the adherents of such views might reasonably expect there to be a causal component to our knowing about external things by means of perception.

Before proceeding to consider each of the two kinds of causal theory, I should issue a cautionary note: the following discussion of causal theories of perception and knowledge will involve a very considerable simplification. Philosophers who have accepted such theories have taken themselves to be theorizing in response to very different questions. Some of them take themselves to be answering the constitutive metaphysical questions that we have just raised,

that is, 'what *is it* for someone to perceive an external thing?' or 'what *is it* for someone to know about that thing by means of perception?' Some philosophers take themselves to be answering weaker modal (but not *constitutive*) analogues of these metaphysical questions, that is, 'what are the necessary and sufficient conditions of someone's perceiving a particular external thing?' and 'what are the necessary and sufficient conditions of someone's knowing about external things by means of perception?' Some philosophers take themselves to be answering the *conceptual* analogues of these questions, that is, 'how do we understand what it is for someone to perceive an external thing?' and 'how do we understand what it is for someone to know about that thing by means of perception?' Some philosophers take themselves to be answering the *semantic* analogues of these questions, that is, 'what are the truth conditions of assertions to the effect that someone perceives an external thing?' and 'what are the truth conditions of assertions to the effect that someone knows about that thing by means of perception?' Some philosophers take themselves to be answering the *pragmatic* analogues of these questions, that is, 'what are we typically doing when we assert that someone perceives an external thing?' and 'what are we typically doing when we assert that someone knows about that thing by means of perception?' In general, the kind of question that a philosopher takes herself to be answering will be influenced by that philosopher's own metaphilosophical views, and this has been especially true in the last century of English-speaking philosophy. This makes the effort of brief survey very complicated. In order to avoid all these complications, I will speak henceforth of some causal relations as being relevant to a *philosophical explanation* of perceiving an external thing, or of knowing about external things by means of perception, and I leave it open what sort of information a philosophical explanation has to provide.

1. What is it for Someone to Perceive an External Thing?

Some philosophers (e.g. Grice 1961; Vision 1997; Audi 1998) have used the phrase 'causal theory of perception' to denote theories that appeal to causation in order to answer the question of what it is for someone to perceive an external thing. According to such 'causal theories of perception', part of what it is for someone to perceive an external thing is for a causal relation of a certain kind to obtain. Different versions of this theory differ in what they take the relata of this causal relation to be. We'll first consider the various kinds of thing that have been claimed to do the relevant sort of causing, and then we'll consider the various kinds of thing that have been claimed to be the relevant kind of effect.

Before proceeding to that survey, we should issue some qualifications: virtually all the proponents of the causal theories to be surveyed in this section will grant that, whenever someone perceives an external thing, there are lots and lots of causal relations that obtain, and these obtain among many different kinds of thing. The point on which our causal theorist insists is that a particular one of these causal relations is part of the correct philosophical explanation of the obtaining of the perceptual relation: it is by virtue of the obtaining of a particular one of these causal relations that the perceptual relation between the person and the external thing obtains. Where causal theories differ is in their answer to the question of *which* of these causal relations it is that has this explanatory significance. (As noted above, whether

this explanatory significance is to be understood in terms of constitution, or in terms of our concepts, or the truth conditions of our ascriptions, or what have you, is a disputed metaphilosophical issue.)

It is commonly thought to be possible for there to be more than one causal relation obtaining among the same two causal relata at a given moment. Might causal theorists agree on what the relata of the relevant causal relation are, and yet disagree on precisely which causal relation between those relata is the explanatorily significant relation? This seems to be a possibility, but it is not a possibility that has been often realized, and that is largely because causal theorists often do no more to specify the explanatorily relevant causal relation than to specify its relata. Many causal theorists are thereby left with some version of the problem of ruling out the 'deviant causal chain', that is, specifying which of the various causal chains that can obtain between the two specified relata is of the right kind to have the relevant kind of explanatory significance. As yet, there is no widely accepted solution to the problem of the deviant causal chain (a problem that arises not just for perception, but also for inference and for action, and that has been more thoroughly discussed among theorists of action than among theorists of perception), and this had led philosophers to wonder whether a causal account of perception can ultimately avoid circularity. Is it possible to specify the kind of causal relation involved in perception, except by specifying it as the kind of causation involved in perception? In my discussion below, I will ignore this complicated issue.

Finally, I will not devote any considerable space to describing the various forms of philosophical opposition to causal theories of what it is for someone to perceive an external thing. That's because there have not been many such opponents—most philosophers who have claimed to oppose causal theories of what it is for someone to perceive an external thing have merely opposed one or another specific causal theory, and have not opposed all such theories in general. Such philosophers might, for instance, claim that, although the obtaining of the right kind of causal relation is a necessary condition of someone's perceiving an external thing, it is not a constituent of the perceiving, or it is not part of our concept of perceiving, or it is not part of a non-circular analysis of the concept of perceiving (for versions of some such criticisms, see Strawson (1974) and Snowdon (1981)), and so on. Thus, they oppose some versions of the kind of causal theory to be discussed in this section, but not all versions. Also, some philosophers (e.g. McDowell 1994) suspect that attempts to provide a theory of perception in entirely non-normative terms are destined to fail, or are motivated by erroneous assumptions, or both. Such philosophers might oppose many specific causal theories, but there is nothing in their complaint that is incompatible with all versions of the causal theory of what it is for someone to perceive an external thing. The causal theory itself does not rule out that perception is fundamentally normative, for the kind of causation that is involved in such perception may itself be normative. Such philosophers need not, therefore, be understood as objecting to the causal theories of what it is for someone to perceive a thing.

Finally, the few philosophers that have opposed all causal theories of what it is for someone to perceive an external thing (e.g. Leibniz, Malebranche) have typically been motivated by one concern only: that they do not countenance any causal interaction between external things, on the one hand, and states of mind, on the other. This concern was perhaps quite natural given Cartesian substance dualism, but it is not as natural, and certainly not as widespread, today.

1.1 What Kind of Thing Does the Relevant Causal Work?

Different causal theorists have held different views about what kind of thing does the relevant causal work in explaining our perceptions of external things. According to Aristotle (1968), Aquinas (1945), Ockham (1957), Reid (1969), Grice (1961), and many others, the relevant causal work is done by the external thing itself. To illustrate: when I see a tomato, these philosophers would say, my seeing is philosophically explained (at least partly) by the fact that the tomato itself is doing a certain kind of causal work. When I hear a tuba, these philosophers would say, my hearing is philosophically explained (at least partly) by the fact that the tuba itself is doing a certain kind of causal work. And when I smell a rose, these philosophers would say, my smelling is philosophically explained (at least partly) by the fact that the rose itself is doing a certain kind of causal work. (Henceforth, all uses of 'explain' will denote philosophical explanation, unless otherwise noted.)

The popularity of some such view throughout the history of Western philosophy may seem natural, since the view itself may seem to be a piece of common sense. But not all philosophers have embraced common sense, and certainly not all philosophers have embraced the particular claim that, when someone perceives an external thing, their perception is explained (at least partly) by *that external thing itself* doing a certain kind of causal work. Thus, according to some philosophers, the relevant causal work is done not by that external thing itself, but rather by its parts. For instance, according to Locke (1975), the relevant causal work is done by corpuscles that make up that external thing. In contrast, other philosophers say that the relevant causal work is done not by that external thing, but rather by something else of which that external thing is a part. For instance, according to Descartes (1984), the relevant causal work is done by extended substance itself (or what we contemporary laypersons might call 'the physical universe'): thus, according to Descartes, for someone to perceive an external thing is, at least in part, for extended substance to cause that person to have certain sensations. According to Berkeley (1975), the relevant causal work is done not by any extended thing, but rather by God, in whom all extended things exist as ideas.

We may graphically represent the range of historically held positions among causal theorists on the issue of what does the relevant kind of causal work in a perceptual episode in the manner of Fig. 27.1.

The origin of the axis in Fig. 27.1 may be taken as the most popular position represented—the position that the relevant causal work is done by the perceived external thing itself. Positions above this origin claim that the relevant causal work is done by something that contains the perceived external thing, whereas positions below the origin claim that the relevant causal work is done by something that is contained in the perceived external thing. Of course, this axis of positions is intended to be nothing more than a visual device—it is not a dimension in any interesting mathematical sense.

God (Berkeley)

Extended substance (Descartes)

External thing (Aristotle, Aquinas, Ockham, Reid, Grice)

Corpuscles (Locke)
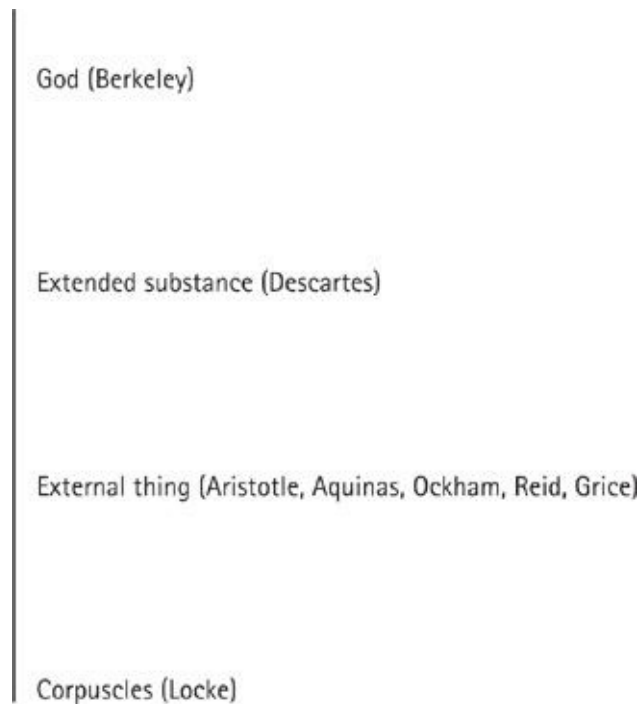
Fig. 27.1

1.2 What Kind of Thing is the Relevant Effect?

Again, different causal theorists have held different views about what kind of thing is the relevant effect in explaining our perceptions of external things.

Descartes (1984), Locke (1975), and Berkeley (1975) all use the term 'idea' to refer to the relevant effect, but they mean different things by 'idea'. For Descartes, an idea is any mental representation, whereas for Locke, I take it, an idea is anything that plays a functional role in our psychological economy. For Berkeley, an idea is any causal impact made upon a spirit. In so distinguishing between the various things that Descartes, Locke, and Berkeley meant by the term 'idea', I leave it open that one or more of these philosophers may have wished to say that mental representations are just those things that play functional roles in our psychological economy, which are in turn just those causal impacts made upon a spirit. But even if they had wished to say this, they would still have disagreed about which of these characterizations is most explanatorily fundamental, and which is derived.

Hume (1978) used the term 'impression' to refer to the effect that is relevant in explaining our perceptions of external things. Given his doctrine of 'no double existence', it may seem odd to describe Hume as having held that our perceptions of external things are explained by causal relations between external things, on the one hand, and impressions, on the other hand. But there is at least one version of such a causal theory that is consistent with the doctrine of 'no double existence': the version that results by taking the external things that we perceive to be ingredients of our impressions of them. Of course, if external things are really to be 'external', then, by definition, it must be possible for them to exist without our having any impression of them. But whenever we *do* have an impression of these external things, then—

according to the version of the causal theory now being envisioned—that impression contains the perceived external thing as an ingredient. It is, we may suppose, possible for the impression to contain this perceived external thing only if the external thing in question exists. And so if the external thing does not exist, then there is no impression that can contain it as an ingredient. Can a cause be an ingredient in the effect that it creates, given Hume's view that cause and effect are distinct existences? That depends upon whether Hume would have thought of parts as distinct from the wholes of which they are parts. Certainly parts are numerically distinct from the wholes of which they are parts, and they are also qualitatively different. But does it follow that they are 'distinct' in Hume's sense of the term? That is not so clear.

Russell (1927) described the effect that is relevant in explaining our perceptions of external things as our acquaintance with our own present percepts, or sense data. Such acquaintance was a relation that we bore to a particular percept, or sense datum. What were these percepts, or sense data, for Russell? That is matter of some scholarly dispute, but Russell clearly thinks that sense data are more fragmentary and transitory than the ordinary objects that common sense would have us perceive, for example, tables, chairs, and speech acts. Two people who, as we ordinarily say, are seeing the same table, or hearing the same speech act, are typically, on Russell's view, acquainted with very different sense data from each other. This is not yet to tell us what sense data are, but it is to tell us something about what they are not.

Grice (1961) describes the effects that are relevant in explaining our perceptions of external things as those states or events that are ascribed to a person when we say, of that person, that it looks (or feels, or sounds, etc.) to her as if p (where p is a schematic variable ranging over clauses—I presume that he has in mind only clauses that are in the indicative mood and present tense, and, perhaps, that include some noun phrase headed by a demonstrative pronoun). The range of verbs that can substitute for 'looks' in this construction will cover the range of familiar perceptual verbs. Grice does not give any further account of the nature of those states or events, except by appeal to the sentences used to ascribe them. For the sake of a label, I will refer to such states and events as 'appearings'.

Armstrong (1961), Pitcher (1971), and Heil (1983) all describe the effects that are relevant in explaining our perceptions of external things as dispositions to believe something or other about the external thing in question. For instance, to perceive a golf ball about 10 feet in front of one is to be disposed to believe something about that particular object, that is, the golf ball. It is not to be disposed to believe that it *is* a golf ball or that it *is* 10 feet in front of one: it is simply to be disposed to believe something or other about it. Of course, one might also be disposed to believe that it is a golf ball or that it is 10 feet in front of one—but that is simply a further fact about one, over and above one's being disposed to believe something or other about that object, and so over and above perceiving that object.

We may graphically represent the range of historically held positions among causal theorists on the issue of what is the relevant kind of effect in a perceptual episode by means of the horizontal axis in Fig. 27.2.

Once again, we may treat the origin of this horizontal axis as the most currently popular position represented on the axis—the position that the relevant effect is the perceptual episode itself. Positions to the right of this origin claim that the relevant effect is something more cognitively sophisticated than the perceptual episode itself, whereas positions to the left of this origin claim that the relevant effect is something less cognitively sophisticated than the

perceptual episode itself. Once again, this axis of positions is intended to be nothing more than a visual device.
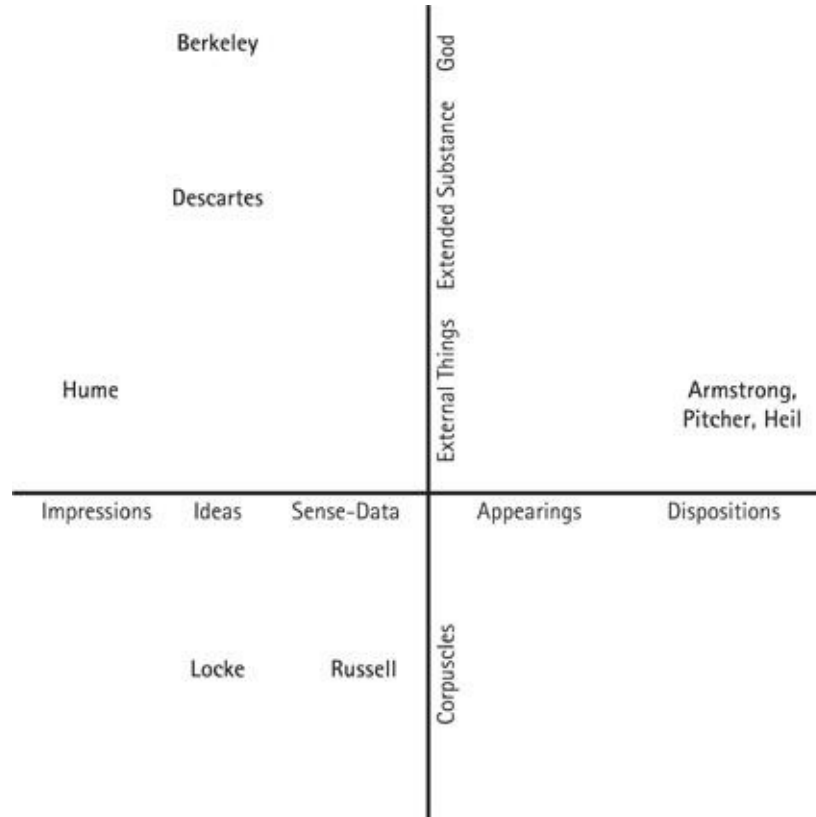
Fig. 27.2 Theories of perceiving an external thing.

By means of the graph at Fig. 27.2, we now depict the variety of theories that appeal to causation in order to explain what it is for someone to perceive an external thing.

2. WHAT IS IT FOR SOMEONE TO KNOW ABOUT EXTERNAL THINGS BY MEANS OF PERCEPTION?

The most common use of the phrase 'causal theory of perception' throughout the last century in the English-speaking philosophical community has been to denote theories that appeal to causation in order to answer the question of what it is for someone to know about external things by means of perception (see e.g. Russell 1927; Price 1932; Alvin Goldman 1967), or as I shall henceforth say, to know about external things 'empirically'. According to such 'causal theories of perception'—which are sometimes called 'causal theories of knowledge'—part of what it is for someone to have empirical knowledge about external things is for a causal relation of a certain kind to obtain. Once again, different versions of this theory differ in what they take the relata of this causal relation to be. We'll first consider the various things that have been claimed to do the relevant sort of causing, and then we'll consider the various things that have been claimed to be the relevant kind of effect. Once again, we will ignore the problem of the deviant causal chain.

Before proceeding however, we should issue two cautionary notes. First, while Alvin Goldman (1967) introduced the phrase 'causal theory of knowledge' to denote a theory that he himself called a 'causal theory of perception', the phrase 'causal theory of knowledge' has typically been used to describe theories of empirical knowledge that are very different from the theories of empirical knowledge that have most typically been called 'causal theories of perception'. Goldman's causal theory of knowledge, for instance, is a version of epistemological *externalism.* According to Goldman's theory and to other such theories, it is *not* a necessary condition of *S's* empirically knowing that *p* is true (where *p* is some truth about external things), that *S* be able to achieve any kind of reflective awareness of a reason for believing that *p* is true. In contrast, most theories of empirical knowledge that have been called 'causal theories of perception' (e.g. Russell 1927; Price 1932) are versions of epistemological *internalism*: they claim that it is a necessary condition of *S's* empirically knowing that *p* is true, that *S* be able to achieve some reflective awareness of reasons for believing that *p* is true. But these latter 'causal' theories add that it is also a necessary condition of *S* empirically knowing that *p* is true (where again, *p* is some truth about external things) that *S* bear some appropriate causal relation to some external things. Thus, although the phrase 'causal theory of knowledge' was originally introduced to denote a version of the theory of empirical knowledge commonly called 'the causal theory of perception', it has now become common to use the two phrases to denote mutually exclusive kinds of theories. Nonetheless, both of these mutually exclusive kinds of theories agree on one point: it is a necessary condition of *S's* empirically knowing that *p* is true (where *p* is some contingent claim about external things) that some appropriate causal relation obtain between (some appropriate feature of) *S* and external things. It is by virtue of their agreement on this one point that I will treat theories of both these mutually exclusive kinds as versions of the causal theory of what it is for a person to know empirically about external things.

Our second cautionary note is this: virtually all philosophers agree that, for *S* to know empirically that *p* (where *p* is some truth about external things), it is necessary that *S* believe that *p*, and that *S*'s belief that *p* be appropriately caused by (or, some philosophers would say, 'properly based upon') some mental state or mental feature of *S*: some of *S*'s beliefs or experiences or memories, or *S*'s rationality, or *S*'s intellectual scrupulousness or virtuousness, or what have you. Thus, virtually all philosophers agree that empirical knowledge of external things requires the obtaining of some causal relation. But, out of fidelity to actual uses of the phrase 'causal theory of perception' or 'causal theory of knowledge', we must not count all such philosophers as causal theorists. Rather, we shall count a philosopher as a causal theorist about our empirical knowledge of external things only if she takes such knowledge to require the obtaining of a causal relation in which the causal work is done by an external thing, and not by some mental state or mental feature of *S*. This policy will still leave us with a generous supply of causal theorists about empirical knowledge, but it will also leave us with a generous supply of non-causal theorists.

2.1 What Kind of Thing Does the Relevant Causal Work?

When someone empirically knows something about external things, what is it that, according to the causal theorist, is doing the relevant causal work? There have been at least

three different answers to this question.

According to Russell (1927), Armstrong (1961), Pitcher (1971), and Heil (1983), it is the external thing that is perceived that is doing the relevant causal work. It is (at least partly) by virtue of that external thing's doing that bit of causal work that the perceiver has empirical knowledge about external things—specifically, about the external things that she perceives.

In contrast, according to Ramsey (1951), Alvin Goldman (1967), and Swain (1981), for S to know empirically that p, the relevant bit of causal work must be done by some fact—typically the fact that p, but in some cases it is a different fact q that is the common cause both of the fact that p and S's belief that p, and it is by virtue of q's doing this causal work that S empirically knows that p.

2.2 What Kind of Thing is the Relevant Effect?

When someone empirically knows something about external things, what is it that, according to the causal theorist, is the relevant effect? Once again, there have been two different answers to this question.

According to Russell (1927), Grice (1961), Harman (1973), Lycan (1988), and Alan Goldman (1988), the relevant effect is one's sensory experience. It is partly by virtue of having a sensory experience of some kind that is appropriately caused that one empirically knows something about external things. In such cases, one's knowledge is empirical because one's knowledgeably held belief is appropriately caused. (For Russell, Grice, and Harman the appropriate causal relation is a relation that the knower believes to hold, whereas for Lycan and Goldman it is not.)
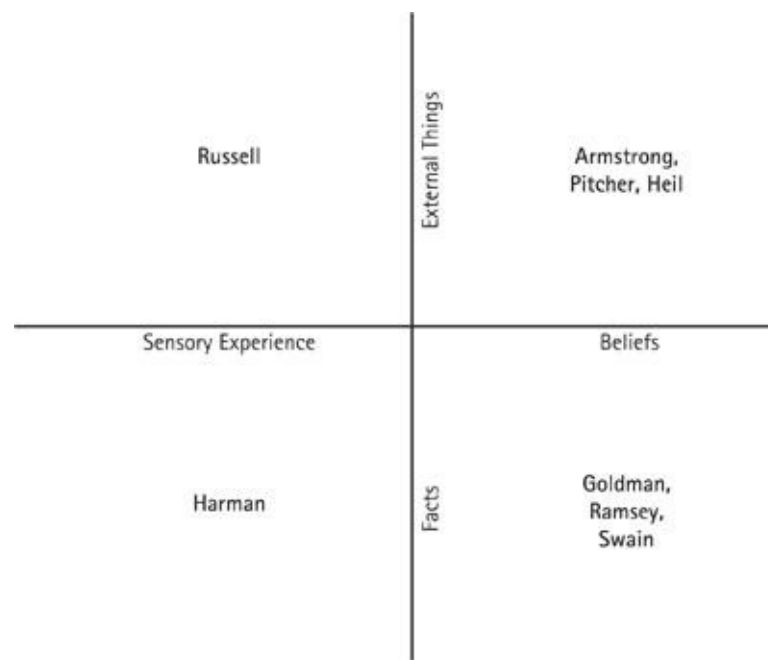
Fig. 27.3 Theories of knowledge by means of perception.

According to Ramsey (1951), Armstrong (1961), Alvin Goldman (1967), Pitcher (1971), Swain (1981), and Heil (1983), the relevant effect is one's belief about external things. It is partly by virtue of having this belief be appropriately caused that one empirically knows something about external things.

We can now map out the various causal theories of empirical knowledge about external things as shown in .

2.3 Recent Disputes Concerning Some Causal Theories of Empirical Knowledge of External Things

What Alvin Goldman (1967) called 'the causal theory of knowledge' was most prominently challenged by Goldman (1976) himself. In that paper, Goldman appeals to an example (of Carl Ginet's) in which a character named 'Henry' sees a real barn in front of him, and is thereby caused to believe that it is a real barn that he sees. But all this does not suffice for Henry to know that it is a real barn that he sees, for Henry is in 'fake barn country', a region in which most of the apparent barns are mere barn façades. What this example shows is that appropriately caused true belief—even appropriately caused justified true belief—does not suffice for empirical knowledge of external things. But this example does not show that some appropriate causal connection is not a constituent—let alone a necessary condition—of empirical knowledge of external things. So this is an example in which something that is commonly regarded as a criticism of a causal theory of knowledge is in fact a criticism of just one version of such a theory. Indeed, many epistemologists today accept that the existence of some causal connection between the fact that p and S's belief that p is a necessary condition of S's empirically knowing that p.

Another version of the causal theory of knowledge—the version most prominently propounded by Harman (1973), according to which S knows that p only if p appropriately causes S's sensory evidence for p, and S abductively infers that p from her sensory evidence—has been widely criticized. Some philosophers, such as van Fraassen (1980), criticize this abductivist causal theory of knowledge by claiming that abductive inference is not a knowledge-transmitting form of inference at all: someone who knows the premises of an abductive inference, and comes to believe the conclusion by performing that inference, cannot thereby gain knowledge of the conclusion. Other critics of Harman's abductivist-causal theory of knowledge allow that abductive inference is knowledge-transmitting, but say that Harman's theory nonetheless cannot account for *all* our empirical knowledge of external things—some of our empirical knowledge of external things cannot be based on abductive inference. Four reasons have been advanced in favour of this latter claim. First, some philosophers, for example, Price (1932), Moore (1968), Chisholm (1989), Lehrer (1990), Alston (1993), Fumerton (1995), and Butcharov (1998) argue that, if propositions about the external world are excluded from our evidence set, then we have no more reason to believe the proposition that our sensory evidence is caused by the impingements upon our sensory organs of physical objects than we do to believe the proposition that our sensory evidence is caused by the intervention of a powerful spirit. The 'spiritual' explanation of our sensory evidence is as well supported as the 'physical object' explanation of our sensory evidence—assuming that the

basis of support is restricted in such a way that no proposition about the external world can itself be included in the supporting basis. Second, some philosophers, for example, Sellars (1963) and Williams (1977), argue that we can't so much as understand the premises of an abductive inference to the existence of external things unless we already have lots of empirical knowledge of external things—again, not *all* of our empirical knowledge about external things can arise from abduction. Third, Strawson (1985) says that, even if all our empirical knowledge of the external world could be arrived at inferentially from our sensory evidence, nonetheless, that is not how most of it was actually arrived at, and so that cannot be the source of its epistemic credentials. And Neta (2004) argues that it is impossible to explain purely subjective facts (i.e. psychological facts that do not imply the existence of any external things) on the basis of facts about external things: there is an explanatory gap between the former and the latter. It remains to be seen whether a causal theorist such as Harman can meet these various challenges.

FURTHER READING

One of the most important contemporary accounts of what it is to see an external thing is provided by Lewis (1980). That account explains what it is for someone to see a thing in terms of counterfactuals relating variations in that person's experience with variations in the thing seen. Those who, like Lewis, hold some version of the counterfactual account of causation would regard Lewis's account of seeing as a causal account. Vision (1997) provides a very comprehensive historical survey of causal theories of what it is for someone to see an external thing. It is widely assumed that philosophical accounts of what it is for someone to see an external thing can be smoothly generalized to the other perceptual modalities.

The seminal work in the causal theory of empirical knowledge is Alvin Goldman (1967). The most influential detailed development of such a theory is in Swain (1981). Philosophers who accept an account of causation on which it involves simply the transmission of a mark will be inclined to regard the theory put forward in Dretske (1981) as a version of the causal theory of empirical knowledge. Philosophers who accept an account of causation on which it involves nothing other than being related by a law of nature may regard the theory put forward in Armstrong (1973) as a version of the causal theory of empirical knowledge. For a comprehensive survey of causal theories of empirical knowledge, see Shope (1983: ch. 5).

REFERENCES

ALSTON, WILLIAM (1993). *The Reliability of Sense Perception*. Ithaca, NY: Cornell University Press.

AQUINAS, THOMAS (1945). *Summa Theologica*, in *Basic Writings of Saint Thomas Aquinas*, ed. Anton C. Pegis. New York: Random House, i.

ARISTOTLE (1968). *De Anima*, trans. D. W. Hamlyn. Oxford: Oxford University Press.

ARMSTRONG, DAVID (1961). *Perception and the Physical World*. London: Routledge & Kegan Paul.

—— (1973). *Belief, Truth, and Knowledge*. Cambridge: Cambridge University Press.

AUDI, ROBERT (1998). *Epistemology: A Contemporary Introduction to the Theory of*

Knowledge. London: Routledge.

BERKELEY, GEORGE (1975). *Philosophical Works*, ed. P. H. Nidditch. London: Everyman's Library.

BUTCHAROV, PANAYOT (1998). *Skepticism about the External World*. Oxford: Oxford University Press.

CHISHOLM, RODERICK (1989). *Theory of Knowledge*, 3rd edn. Englewood Cliffs, NJ: Prentice-Hall.

DESCARTES, RENÉ (1984). *The Philosophical Writings of Descartes*, trans. John Cottingham, Robert Stoothoff, and Dugald Murdoch. Cambridge: Cambridge University Press, ii.

DRETSKE, FRED (1981). *Knowledge and the Flow of Information*. Cambridge, Mass.: Bradford Books.

FUMERTON, RICHARD (1995). *Metaepistemology and Skepticism*. Lanham, Md.: Rowman & Littlefield.

GOLDMAN, ALAN (1988). *Empirical Knowledge*. Berkeley and Los Angeles: University of California Press.

GOLDMAN, ALVIN (1967). 'A Causal Theory of Knowing', *Journal of Philosophy* 64: 355–72.

—— (1976). 'Discrimination and Perceptual Knowledge', *Journal of Philosophy* 73: 771–91.

GRICE, H. P. (1961). 'The Causal Theory of Perception', *Aristotelian Society Proceedings, Suppl*. 35: 121–52.

HARMAN, GILBERT (1973). *Thought*. Princeton, NJ: Princeton University Press.

HEIL, JOHN (1983). *Perception and Cognition*. Berkeley and Los Angeles: University of California Press.

HUME, DAVID (1978). *A Treatise of Human Nature*, ed. P. H. Nidditch. Oxford: Clarendon.

LEHRER, KEITH (1990). *Theory of Knowledge*. Boulder: Westview.

LEWIS, DAVID (1980). 'Veridical Hallucination and Prosthetic Vision'. *Australasian Journal of Philosophy* 58: 239–49.

LOCKE, JOHN (1975). *An Essay Concerning Human Understanding*, ed. P. H. Nidditch. Oxford: Oxford University Press.

LYCAN, WILLIAM G. (1988). *Judgment and Justification*. Cambridge: Cambridge University Press.

MCDOWELL, JOHN (1994). *Mind and World*. Cambridge, Mass.: Harvard University Press.

MOORE, G. E. (1968). 'The Status of Sense- Data', in *Philosophical Studies*. Totowa, NJ: Littlefield, Adams.

NETA, RAM (2004). 'Skepticism, Abductivism and the Explanatory Gap', in *Philosophical Issues: A Supplement to Noûs*, ed. Ernest Sosa and Enrique Villanueva. Boston: Blackwell.

OCKHAM, WILLIAM (1957). *Quodlibeta* in *Philosophical Writings*, trans. Philotheus Boehner. London: Thomas Nelson & Sons.

PITCHER, GEORGE (1971). *A Theory of Perception*. Princeton, NJ: Princeton University Press.

PRICE, H. H. (1932). *Perception*. London: Methuen.

RAMSEY, F. P. (1951). *The Foundations of Mathematics, and other Logical Essays*. London: Routledge.

REID, THOMAS (1969). *Essays on the Intellectual Powers of Man*. Cambridge, Mass.: MIT.

RUSSELL, BERTRAND (1927). *The Analysis of Matter*. New York: Dover.

SELLARS, WILFRID (1963). 'Phenomenalism', in *Science, Perception and Reality*. Atascadero, Calif.: Ridgeview.

SHOPE, ROBERT (1983). *The Analysis of Knowing: A Decade of Research*. Princeton, NJ: Princeton University Press.

SNOWDON, PAUL (1981). 'Perception, Vision and Causation', *Proceedings of the Aristotelian Society* 81: 175–92.

STRAWSON, P. F. (1974). 'Causation in Perception', *Freedom and Resentment*. London: Methuen, 66–84.

—— (1985). *Skepticism and Naturalism: Some Varieties*. New York: Columbia University Press.

SWAIN, MARSHALL (1981). *Reasons and Knowledge*. Ithaca, NY: Cornell University Press.

VAN FRAASSEN, BAS (1980). *The Scientific Image*. Oxford: Oxford University Press.

VISION, GERALD (1997). *Problems of Vision*. Oxford: Oxford University Press.

WILLIAMS, MICHAEL (1977). *Groundless Belief: An Essay on the Possibility of Epistemology*. New Haven: Yale University Press.

CHAPTER 28
CAUSATION AND SEMANTIC CONTENT

FRANK JACKSON

1. INTRODUCTION

Thought is directed towards the world. I believe that it will rain soon; I desire peace in the Middle East; I hope for long life. These thoughts are about the world. More particularly, my belief is about how things have to be for it to be true, and my hope and desire are about how things have to be for them to be satisfied. Further, it is a commonplace that beliefs may be false and desires and hopes unfulfilled. This is the sense in which they can be about what does not exist, in which they have intentionality, as it is often put. When you believe falsely that the cheque is in the mail, your belief is about a cheque in the mail in the sense that in order for your belief to be true, there has to be a cheque in the mail.

How does causation enter the picture? Belief is a state shaped by the world, a state that seeks to fit the world; desire is a state that shapes the world, that seeks to make the world fit it. Both metaphors are compelling and are loaded with causality. Thus the long tradition of seeking to understand the aboutness of belief and desire, and intentional states in general (thoughts, for short), in causal terms. We often use 'reference' for the relation between thought and world. We often use 'content' for how things have to be for, for example, a belief with that content to be true and a desire with that content satisfied. In these terms, the tradition of seeking to understand aboutness in causal terms is the tradition of seeking causal accounts of reference and content.

We humans make public the contents of our thoughts using sentences. I gave the content of the belief that there is a cheque in the mail using the sentence 'the cheque is in the mail'. The topic of the content and reference of thought and the topic of the content and reference of language dovetail. How the dovetailing goes depends on matters to be addressed later. I start with causal approaches to the content and reference of thought.

2. STARTING WITH THE EPISTEMOLOGY

There is no great mystery about how we arrive at views about what agents believe and desire. We infer them from the way subjects interact with their environments. Someone who buys red wine and not white wine is inferred to desire red wine more than white wine. Someone who reaches for an umbrella is inferred to believe that it will rain. More precisely, we form our hypotheses about what someone believes and desires by balancing off hypotheses about what's believed and what's desired. Inferring the belief that it will rain from reaching

for an umbrella only makes good sense on the assumption that the agent desires to remain dry. Buying red wine is good evidence for a preference for red wine on the assumption that the agent believes that he will be consuming it. Roughly, the way we balance belief against desire is by reference to what is sometimes called the axiom of belief-desire psychology: subjects behave in such a way that if what they believe is true, they get what they desire. In the jargon: in arriving at hypotheses about what subjects believe and desire, we look for the beliefs and desires that would *rationalize* their behaviour. This is rough indeed, neglecting as it does the important place of strength of belief and desire. Someone may believe that Hyperion will win the cup and desire money. It doesn't follow that they'll bet on Hyperion. They may judge that the odds on Hyperion are too short and that Hydrogen's odds are a bargain and put their money on Hydrogen. All the same, the rough sketch will serve our needs.

Suppose you knew everything there was to know about how some subject behaves and would behave in all possible circumstances, and suppose you are the complete master of the use of the axiom of belief-desire psychology to arrive at hypotheses about belief and desire going by behaviour. Call your hypothesis about what our subject believes and desires, the ideal rationalization of the subject's behaviour. One theory about what a subject believes and desires says that it is given by the ideal rationalization of their behaviour. This may sound like a subjective or interpretationist account of the content of belief and desire with its reference to your hypothesis, but it isn't really. The real work is being done by the stipulation that you have *full* information and *complete* mastery. You entered the account as a rhetorical device only.

Why not accept the ideal rationalization account of the content of belief and desire? For two reasons: one is that we have a classic case of underdetermination. We have one body of data: behaviour in circumstances, required to determine the answer to two unknowns: what's believed and what's desired. An action rationalized by the desire to win money and a belief that stock S will rise can equally be rationalized by a desire to lose money and a belief that S will drop. We need more ammunition.

The second reason relates to the debate over narrow versus wide content. The ideal rationalization account of content is wide in one good sense. It is all to do with making sense of subjects' behaviour in the *context* of their environment, their surroundings. It is wide in the same sense as being water-soluble is wide. To be water soluble is, roughly, to be such as to dissolve when placed in water. It concerns how a substance is *related* to water. However there is a sense in which being water-soluble isn't wide. Two substances alike in themselves are alike in whether or not they are water-soluble. This is because what matters for being water-soluble is the totality of actual and possible interactions with water, and that totality is the same for any two substances alike in their internal make-up. However, many hold that two agents alike in themselves may have thoughts with different contents. Content does not supervene on internal nature. It is in part a function of the environment a subject finds herself in.

Suppose, for example, we thought of belief on the model of internal states which carry information about their surroundings by virtue of causally co-varying with some feature of it. A petrol gauge carries information about the level of gas left in the tank by virtue of the way its pointer's position causally co-varies with the level of gas left. But the information it carries

depends on how it is connected up to the tank and on the tank it is in fact connected to. You could 'reverse wire' it so that it indicated the gas consumed instead of the gas left, and it could be connected to any of a number of tanks, or the gauge could be connected to the radiator so that it carried information about the water in the radiator, and so on. If anything like this is the right way to think of belief, the content of belief is wide in the sense of being a function of environment as well as internal nature.

Or suppose we thought of belief in the terms suggested by evolutionary theory. The survival value of having states that co-vary with one's surroundings is obvious. We need good predator detectors and good food detectors. So we might think of belief that p as the state selected to co-vary with p. Likewise, the survival value of having states that make certain changes to the environment is obvious. We need states that take us away from predators and near to food. So we might think of desire that p as the state selected to bring about p. Again we would have an approach to content that made it wide. Content would depend on selectional history as well as internal nature.

So informational and selectional approaches to mental content both differ from the ideal rationalization of behaviour approach in making content wide. They also promise to assist with the underdetermination problem that besets the ideal rationalization approach. Looking to what a state co-varies with and to what it was selected for promises to reduce the number of possible assignments of content. It cuts down the range from which we need to make our choice.

All the same, there are serious problems for both the informational and selectional theories, as we will now see.

3. INFORMATIONAL TREATMENTS OF MENTAL CONTENT

Informational or indicator approaches to content can be developed in many ways. One popular way draws on a view many hold to be independently plausible, namely, that thoughts are sentence-like structures in the brain, the famous language of thought hypothesis. The thought that a is F is a structure in the head made of a mental word for a, a^m, and a mental word for F, F^m. The thought counts as a belief if the sentence in *mentalese* does belief-like things—responds to the environment in ways designed to fit the facts; it counts as a desire if the sentence in mentalese does desire-like things—changes the environment in ways designed to fit it. The content of the thought is given by appeal to an informational treatment of the reference of a^m and of F^m. Soif a^m co-varies with the presence of cats, a^m stands for cats; if F^m co-varies with being furry, F^m stands for being furry; and the thought's content will be that cats are furry.

Everyone agrees that this is an interesting, attractively simple approach to the content of mental states. They also agree that it has many problems and the (increasingly complex) debate is over whether or not it can get around them. Here's a sample of the problems. (1) The view cannot be that a^m stands for cats if and only if a^m's presence in the brain co-varies with cats. That would make it impossible to think that cats are furry in the absence of cats. It won't

do to say that all that's required is that often a^m co-varies with cats. It is possible to think that cats are furry while living most of one's life on a desert island where there are no cats. It won't do to say that all that is needed is a *learning period* where a^m co-varies with cats and gets to refer to cats ever after, so allowing our person on the desert island to think that cats are furry provided there was a learning period during which the co-variation obtained. For obvious reasons, the co-variation is often worst during the learning period—it is a *learning* period afterall. (2) The local nature of causation means that co-variation with items in the world around us is underpinned by co-variation with elements of the causal chain from those items to our brains. To the extent that there is co-variation between some token in our heads and cats, there will be co-variation between, for example, what's happening at our peripheries in the causal path from cats to that token in our heads. The question for the informational approach is to say what makes it the case that the token is about cats and not the peripheral sensory stimulation pattern that results from contact with cats. One way to reinforce the point is to note that a great many of our thoughts are about things we never come across. We read about them in books. My thoughts about many of the creatures on the Galapagos Islands do not come from co-variation between what's in my head and the presence of those creatures. At best, there is a kind of co-variation at a distance. Perhaps those who write the books have thoughts about the lizards by virtue of a co-variation between their head states and the lizards. My head states then co-vary with the presence of words on the pages of the books and so in turn with the states of those who wrote the books. But by far the most reliable co-variation in my case will be between words and head states, how in that case do I get to have thoughts about the *lizards*?

4. SELECTIONAL TREATMENTS OF MENTAL CONTENT

The proposal that belief that *p* is the state selected to co-vary with *p*, and that desire that *p* is the state selected to bring about *p*, is a rough sketch to give the general idea. It is often called a teleological theory of content, on the ground that selection delivers the purpose or *telos* of a state.

An obvious question to ask of it is how it might handle beliefs and desires about matters too recent to play a selectional role. Being bad at handling cars reduces one's prospects of surviving long enough to have children but this has not been the case for long enough to have selectional effects. All the same, we have thoughts about cars. The usual way of tackling this problem is to think of thoughts about cars as built out of more basic elements that were themselves selected for. I will, though, focus on two problems of a general kind that do not turn on the detail of one or another way of spelling out the selectional or teleological theory. They are problems we can raise without moving beyond our rough sketch.

The first is the famous swampman problem. It is possible, although *extremely* unlikely, that a collection of molecules from a swamp should rearrange itself to make an exact duplicate of, say, me. This duplicate would look exactly like me and would act just as I do in any given situation. There is a strong intuition that it would have beliefs and desires. It would certainly be indistinguishable, internally and externally, from something having beliefs and desires,

namely, me. If the selectional theory is correct, swampman does not have beliefs and desires: its thoughts are not about anything because they were not selected for anything. Some teleologists bite this bullet but attempt to soften the blow by allowing that swampman would have *sensory* states just like mine. But they would be very curious sensory states. They would have to be pains that swampman did not desire to cease and were not accompanied by the belief that something untoward was happening, and taste sensations that could not evoke desires that they continue. Some teleologists spit the bullet out by arguing that swampman would have contentful thoughts but in a secondary or derived sense. Swampman's states would be like mine and mine have the right selectional history to have content. The trouble with this suggestion is that they are *not* like mine in the sense that matters according to the selectional theory.

The second problem comes from the coherence of creationism. We might discover, to our very great surprise, that creationism is true. But that would not be the discovery that we had no beliefs and desires. Some teleologists insist that there is no problem here. We would in that case be God's creation and our thoughts would be designed by him to have the contents that they do. (These theorists insist that their theory is a *design* account of content and that the selectional story enters the picture because, as a matter of fact, evolutionary selection is the way design is done in our world.) But where do the contents of God's thoughts come from? God is a necessary being. His or her states were not selected for, and nor were they designed.

5. Linguistic Content

We said that the content of thought and the content of language dovetail, but how? A very attractive answer is by virtue of language users entering into agreements to convey what they think using words. I believe that the object in front of me is a circle. I want to tell you this. The word 'circle' is a good word for me to use. We might have settled on the word 'square' to do this job but we didn't. The suggestion is not that we entered into explicit agreements to use words as we do, including using the word 'circle' for circles. Scientists do sometimes have conferences to thrash out an agreement on how to use a term—a recent example is the word 'planet'—but mostly the agreements are implicit ones we pick up as we acquire mastery of a language. This means that the question of the content of some expression in some natural language is the question as to what we agreed to use that word to say about how we take things to be. This makes it a very different question from the question as to what some word in mentalese means. We did not enter into agreements of any kind to use the words of mentalese in one way rather than in another.

What then did we agree to use various words for? In some cases the answer seems relatively easy. I chose the example of the word 'circle' because I thought that there would be little controversy about what it stood for. It stands for this shape: O. Other cases are harder and prominent among the hard cases are proper names. The rest of this chapter will be concerned with them.

6. The Description Theory of Proper Names

The circumstances under which we use proper names suggest that their reference goes by description. Consider philosophy students attending their first lecture and being told, as they might be, that there was someone called 'Aristotle', that he lived a long time ago, that he devised an important system of logic, and that in next week's lecture they will learn much more about him. They produce the sentence 'Aristotle will be the subject of next week's lecture' as a claim about how they take things to be. They would obviously be justified in doing this, but what else have they learnt other than that someone satisfying various descriptions that they've learnt about in the first lecture will be the subject of next week's lecture?

Equally, the way we come to identify the referent of a name suggests that its reference goes by description. I come to believe that the lecture I am attending is given by Hilary Putnam, that is, that 'Hilary Putnam' refers to the person giving the lecture, by virtue of information of the following kind: that that name appeared on the lecture poster, someone I trust said 'Hilary Putnam is the lecturer', the lecturer looks like the photograph on the back of a book bearing the name 'Hilary Putnam' on its spine, and so on. I use descriptive information to settle that 'Hilary Putnam' refers to a certain person. Indeed it is hard to see what other kind of information I could reasonably use.

Nevertheless, the description theory has been subjected to powerful criticism, notably by Saul Kripke, and many, possibly most, philosophers of reference nowadays prefer some kind of causal-historical theory of reference for proper names. The rest of this chapter will review the debate between description and causal-historical theories of reference for proper names.

7. KRIPKE'S THREE-PRONGED ATTACK ON THE DESCRIPTION THEORY

7.1 First Prong

There is a relatively uncontroversial sense in which the reference of names goes by descriptions. Reference supervenes on nature. If name 'N' refers to x and not to y, then x and y must differ in some way over and above the difference in regard to reference. The live debate is over whether or not, when 'N' refers to x in speaker S's mouth, S *associates* with 'N' a description or set of descriptions that applies to x alone. The debate is over the availability to speakers of descriptions that are uniquely true of the referents. The first argument—sometimes known as the objection from error and ignorance—against the description theory is that there are cases where a name refers to x in users' mouths and yet many of the users' cannot cite a description uniquely true of x. Take our beginning philosophy students. It is very plausible that 'Aristotle' refers to Aristotle in their mouths but, so the argument goes, they do not know of descriptions uniquely true of Aristotle: the name is not that unusual; anyway he wasn't called 'Aristotle' but by the corresponding name in ancient Greek and the students will not know what that name was; and, finally, the students don't know enough about his doctrines after one lecture to pick him out from a whole range of ancient philosophers.

How then does it come to be that 'Aristotle' refers to Aristotle? Opponents of the description theory argue, surely plausibly, that we know the answer in rough outline. In the

past, Aristotle was given the corresponding Greek name. This set up a practice of using the name in sentences to pass on information about him, from one user of the name to the next, ending up with the uses of the lecturer and the students in our example. We have a causal-historical information-preserving chain of reference borrowing if you like—one that allows that the configuration of the name will change as the chain passes through different language communities—and that's how reference is secured on the causal-historical theory of reference for names.

The preceding paragraph is a very rough sketch of the causal-historical theory but it is enough to raise a serious problem for the error-ignorance objection to the description theory. The story about the name 'Aristotle' is common knowledge. It is not something the students will be unaware of. Here it is important to bear in mind that we can name objects with which we have no causal contact. If astro-physicists conclude that there is an especially interesting star outside the light cone and name it 'Hero', it would be misguided for philosophers to tell them they couldn't. The causal-historical theory is, that is, only plausible for some names. Which names? The obvious answer is that it is good for the ones we decided to use in the relevant way. Attaching names to items and then using these baptisms to underpin information-preserving causal chains is something we often do with names. But this is common knowledge. The way we use sentences containing proper names in finding our way around our world, and as a source of knowledge about it, shows that we know about naming practices and about how they generate information-preserving chains. This is how we know that the appearance of 'Real Madrid won last night' in the morning paper gives us information about a happening spatially and temporally distant from the token sentence. The serious problem for the error-ignorance objection to the description theory is that it is common knowledge that a causal-historical theory is right for many names. The students do know a description that uniquely picks out Aristotle. Aristotle alone is the right kind of causal origin of token sentences containing 'Aristotle' that carry information about Aristotle.

I do not mean that the students could give a neat statement of the relevant description. They couldn't. But if they watch a programme on TV concerned to identify Aristotle, they'd know which descriptions were the right ones and which the wrong ones to settle who Aristotle was. Neither they nor the makers of the programme would need to consult philosophers of reference. Folk knowledge would be enough.

7.2 Second Prong

The second objection is often called the modal objection. Normally definite descriptions are non-rigid in the sense that their references can vary under counterfactual hypotheses. 'The 43rd President' of the United States refers to George W. Bush, but had he lost the 2004 election, it would have referred to John F. Kerry. However 'George W. Bush' and 'John F. Kerry' always refer to the very same people. Names are rigid; typically definite descriptions are not.

Description theorists take the point and modify the description theory. They argue that names should be thought of as *rigidified* definite descriptions. The reference of a definite description is given by the rule: 'the D' refers at world w to the unique thing that is D in w if

such there be. The reference of a rigidified definite description (where we use 'actual' to secure the rigidification) is given by the rule: 'the actual *D*' refers at world *w* to the unique thing that is *D* in the actual world if such there be. Thus the reference of 'the actual *D*' does not vary as we go from possible world to possible world.

7.3 Third Prong

The third objection to the description theory is often called the epistemic objection. If a name '*N*' is equivalent to 'the *D*', for some *D*, we could not discover that *N* lacked *D*. That would be discovering that the *D* wasn't *D*! The point stands if we replace 'the *D*' by 'the actual *D*'. One couldn't discover that the actual winner of the 2004 election wasn't the winner of the 2004 election, anymore than one could discover that the winner wasn't the winner. In other words, description theorists cannot meet the epistemic objection by going rigid; this is why it is the objection of choice for many opponents of the description theory.

It is certainly true that one might discover that many of the things one took to be true of George W. Bush were not in fact true of him. One might discover that he was not the 43rd President—amazingly there's been a counting error that has escaped notice until now. He was in fact the 44th President. One might discover that his wife's name is not Laura—she recently changed her name and the news hasn't got out yet. One might discover that he isn't a Republican—he's been a secret, very secret, Democrat all these years. But is it true that, for every description one takes to be true of Bush, one might discover that it is not true of him? The answer to this question seems to be a clear no. For there is such a thing as discovering that someone one took to be George W. Bush is not in fact George W. Bush; mistaken identity is a commonplace. And when one discovers that someone is not George W. Bush, one does so by finding that they have properties that precludes their being Bush. I know, for example, that the person sitting next to me on the train is not George Bush because I know he has the wrong properties to be Bush. There are Bush-precluding properties and he's got some of them. Some object that I could discover that the person next to me on the train is Bush. Imagine that I ask him who he is and get answers supporting his being Bush; imagine he takes off a face mask I hadn't noticed and he looks just like Bush; imagine that the train is met by a host of reporters; etc. But that's to *change* the properties of the person next to me on the train—I know, for instance, the person next to me won't be met by reporters and isn't wearing a face mask. We can all agree that changing enough properties of *x* could make it reasonable to hold that *x* is George W. Bush. The issue was, are there properties that preclude being Bush?—to which the answer is yes—not, could we find properties to support being Bush?

8. SYMMETRICAL WORLDS AND THE DESCRIPTION THEORY

I finish with an objection that has become prominent since Kripke's discussion. Imagine that physicists establish that our world divides into two widely separated regions that are exact duplicates of each other, and that I know this. For ease of reference, let us suppose that one region is known to be earlier than the other in time. And suppose that I cannot tell, and know that I cannot tell, if I am the early Jackson or the late Jackson. In this case I will be unable to

tell if I refer to the early or the late London, when I use 'London' in the sentence 'The conference is in London.' All the same, I will be referring to either the early or to the late London.

An example of this kind is sometimes cited as a killer objection to the description theory of names. But what exactly is the objection? The point seems to be that early and late Jackson must associate the very same descriptions with names such as 'London' but, runs the objection, 'London' in early Jackson's mouth refers differently from 'London' in late Jackson's mouth. But when we are dealing with names such as 'London'—names whose job is to provide information through the way they figure in baptism-generated causal chains—the plausible descriptions for a description theorist to invoke will involve *relations to tokens*: that's how the tokens get to provide the information. A petrol gauge provides information about happenings that stand in so and so relation to it. This means that it makes perfect sense from the description point of view that the early tokens of 'London' and the late tokens of 'London' should differ in reference. The difference in reference will be the product of the fact that the same descriptive relation holds between two *different* sets of tokens of 'London'—the early set and the late set—and the two different cities.

Further reading

Braddon-Mitchell and Jackson (2006) discuss some of the issues of this chapter in ways congruent with the approach taken here (for the obvious reason). There is a glossary at the end with definitions of, for example, belief-desire psychology, intentionality, wide content, the language of thought, etc. Dretske (1988) is a good source for informational approaches to mental content and the topic of explaining behaviour. Papineau (1993) is a good source for teleological treatments of content; so is the more difficult Millikan (1993). Kripke (1980) is the classic text for the attack on the description theory of reference; see also Putnam (1975) and recently Soames (2002). Putnam (1975) is an important impetus for the thesis that content, both mental and linguistic, is wide. For a defence of the description theory, see Jackson (1998) and Searle (1983). A helpful discussion that links the issues of intentionality, theories of reference for names, and the debate over wide content is Segal (2005).

References

BRADDON-MITCHELL, DAVID, and JACKSON, FRANK (2006). *The Philosophy of Mind and Cognition*. 2nd edn. Oxford: Blackwell.

DRETSKE, FRED (1988). *Explaining Behavior*. Cambridge, Mass.: MIT.

JACKSON, FRANK (1998). 'Reference and Description Revisited', *Philosophical Perspectives*, xii. *Language, Mind, and Ontology*, ed. James E Tomberlin. Cambridge, Mass.: Blackwell, 201–18.

KRIPKE, SAUL (1980). *Naming and Necessity*. Oxford: Blackwell.

MILLIKAN, RUTH (1993). *White Queen Psychology and Other Essays for Alice*. Cambridge, Mass.: MIT.

PAPINEAU, DAVID (1993). *Philosophical Naturalism*. Oxford: Blackwell.

PUTNAM, HILARY (1975). 'The Meaning of "Meaning" ', in *Mind, Language and Reality*. Cambridge: Cambridge University Press, 215–71.

SEARLE, JOHN (1983). *Intentionality*. Cambridge: Cambridge University Press.

SEGAL, GABRIEL (2005). 'Intentionality', in Frank Jackson and Michael Smith (eds.), *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press, 283–309.

SOAMES, SCOTT (2002). *Beyond Rigidity*. New York: Oxford University Press.

CHAPTER 29
CAUSATION AND EXPLANATION

PETER LIPTON

1. INTRODUCTION

There are intimate connections between causation and explanation, between cause and because, and this suggests the projects of illuminating one in terms of the other. Aristotle, for example, appears to have understood his notion of efficient causation in terms of explanation (Sorabji 1980). Nowadays it is not so common to attempt to use explanation to analyse the metaphysics of causation, though there is considerable interest in using explanation to analyse the *epistemology* of causation, to see explanatory considerations as a guide to causal inference (Lipton 2004). On the metaphysical level, the analysis has more commonly gone in the other direction, to causal theories of explanation rather than explanatory theories of causation (Salmon 1984; Lewis 1986; Woodward 2003). As readers of this *Handbook* will be aware, the nature of causation is itself highly contested, but the absence of an agreed account has not stopped philosophers from helping themselves to the notion of causation to account for other things, and philosophers of explanation are no exception. The practice is benign and potentially illuminating, so long as one does not go on to attempt to analyse causation in terms of explanation and so create a circle.

In its simplest form, a causal model of explanation maintains that to explain some phenomenon is to give some information about its causes. This prompts four questions that will structure the discussion to follow. The first is whether all explanations are causal. The second is whether all causes are explanatory. The answer to both of these questions turns out to be negative, and seeing why this is so helps to clarify the relationship between causation and explanation. The third question is itself a request for an explanation: Why do causes explain, when they do? Why, for example, do causes explain their effects but effects not explain their causes? Finally, we will consider how explanation can illuminate the process of causal inference.

2. ARE ALL EXPLANATIONS CAUSAL?

The causal model of explanation has considerable attractions. Both science and ordinary life are filled with causal explanations, and the causes we cite seem explanatory precisely because they are causes. Indeed it appears that requests for explanation, why-questions, can often be paraphrased as what-is-the-cause-of questions. Moreover, the causal model passes three key tests for any adequate account of explanation, tests that other popular models of

explanation fail. The first of these is that a model should account for the difference between knowing and understanding. Knowing that something is the case is necessary but not sufficient for understanding why it is the case. We all know that the sky is sometimes blue, but few of us understand why. Typically, when people ask questions of the form 'Why *P*?', they already know that *P*, so understanding why must require something more than knowing that. The causal model gives a natural account of this gap, since we can know that something occurred without knowing what caused it to occur. (By contrast, a model according to which we understand why something occurs by seeing that it was to be expected (Hempel 1965: 337, 364–76) seems not to pass this test, since simply knowing that *P* will often already involve having good reasons to believe that *P* and indeed good reasons to expect *P*.)

The second test is the test of the why-regress. As most of us discovered in our youth and to our parents' consternation, whatever answer someone gives to a why-question, it is almost always possible sensibly to ask why the explanation itself is so. Thus there is a potential regress of explanations. If your daughter asks you why the same side of the moon always faces the earth, you may reply that this is because the period of the moon's orbit around the earth is the same as the period of the moon's spin about its own axis. This may be a good explanation, but it does not preclude your daughter from going on to ask the different but excellent question as to why these periods should be the same. For our purposes, the salient feature of the why-regress is that it is benign: the answer to one why-question may be explanatory and provide understanding even if we have no answer to why-questions further up the ladder. This shows that understanding is not like some substance that gets transmitted from explanation to what is explained, since the explanation can bring us to understand why what is explained is so even though we do not also understand why the explanation itself is so. Any account of understanding that would require that we can only use explanations that have themselves been explained fails the test of the why-regress. The causal model passes this test, because it is possible to know that *C* caused *E* without also knowing what caused *C*. The model thus shows why the why-regress is benign. (By contrast, a model according to which we explain a phenomenon by reducing it to something familiar (cf. Friedman 1974: 9–11) may not account so well for the why-regress, since if the familiarity model were correct, then the explanation would presumably not itself support a why-question, since it is already familiar. But almost all explanations do support a further why-question.)

The third test is the test of self-evidencing explanations (cf. Hempel 1965:370–4). These are explanations where what is explained provides an essential part of the reason for believing that the explanation itself is correct. Self-evidencing explanations are common, in part because we often infer that a hypothesis is correct precisely because it would, if correct, provide a good explanation of the evidence. Seeing the disembowelled teddy bear on the floor, with its stuffing strewn across the living room, I infer that Rex has misbehaved again. Rex's actions provide an excellent if discouraging explanation of the scene before me, and this is so even though that scene is my only direct evidence that the misbehaviour took place. To take a more scientific and less destructive example, the velocity of recession of a galaxy explains the red shift of its characteristic spectrum, even if the observation of that shift is an essential part of the scientist's evidence that the galaxy is indeed receding at the specified velocity. Self-evidencing explanations exhibit a kind of circularity: *H* explains *E* while *E* is evidence for *H*.

As with the why-regress, however, what is salient is that there is nothing vicious here: self-evidencing explanations may be illuminating and well supported. Any account of understanding that rules them out is incorrect. The causal model passes this test too. It allows for self-evidencing explanations, because it is possible for C to be a cause of E where knowledge of E is an essential part of one's reason for believing that C is indeed a cause. (By contrast, a rational expectation model seems to fail this test too, since if A explains B by giving a reason to believe B, then to suppose that B simultaneously gives a reason to believe A would be to move in a vicious circle; at least it cannot be that A is my reason for B and B is my reason for A.)

Alongside all these virtues, however, the causal model of explanation has an obvious limitation, because not all explanations are causal. Mathematicians and philosophers, for example, give explanations, but mathematical explanations are never causal, and philosophical explanations seldom are. A mathematician may explain why Gödel's Theorem is true, and a philosopher may explain why there can be no inductive justification of induction, but these are not explanations that cite causes. There are even physical explanations that seem non-causal. Here are two striking examples. First, suppose that a bunch of sticks is thrown into the air with a lot of spin, so that the sticks separate and tumble as they fall. Now freeze the scene at a moment during the sticks' descent. Why are appreciably more of them near the horizontal axis than near the vertical, rather than in more or less equal numbers near each orientation one might have expected? The answer, roughly speaking, is that there are many more ways for a stick to be near the horizontal than near the vertical. To see this, consider purely horizontal and vertical orientations for a single stick with a fixed midpoint. There are indefinitely many horizontal orientations, but only two vertical orientations. Or think of the shell that the ends of that stick trace as it takes every possible orientation. The areas that correspond to near the vertical are caps centred on the north and south poles formed when the stick is 45° or less off the vertical, and this area is substantially less than half the surface area of the entire sphere. Another way of putting it is that the explanation why more sticks are near the vertical than near the horizontal is that there are two horizontal dimensions but only one vertical one. This is a lovely explanation, but apparently not a causal explanation, since geometrical facts cannot be causes.

The second example of a non-causal explanation concerns reward and punishment (Kahneman, Slovic, and Tversky 1982: 66–8). Air Force flight instructors had a policy of strongly praising trainee pilots after an unusually good performance and strongly criticizing them after an unusually weak performance. What they found is that trainees tended to improve after a poor performance and criticism; but they actually tended to do worse after good performance and praise. What explains this pattern? Perhaps it is that criticism is much more effective that praise. That would be a causal explanation. But the pattern of performance is also what one should expect if neither praise nor criticism had any effect. It may just be regression to the mean: extreme performances tend to be followed by less extreme performances. If this is what is going on, we can explain the observed pattern by appeal to chance rather than to any cause. (This example ought to give pause to parents who are too quick to infer that punishing children for bad behaviour is more effective than rewarding them for good behaviour.)

The existence of non-causal explanations shows that a causal model of explanation cannot

be complete. One reaction to this would be to attempt to expand the notion of causation to some broader notion of 'determination' that would encompass the non-causal cases (Ruben 1990: 230–3). This approach has merit, but it will be difficult to come up with a such a notion that we understand even as well as we understand causation, without falling into the relation of deductive determination, which will expose the model to many of the objections to the deductive-nomological model, according to which an explanation is a valid argument whose conclusion is a description of the phenomenon to be explained and whose premises include essentially at least one law (Hempel 1965: ch. 10). That model faces diverse counterexamples of deductions that are not explanatory, such as the deduction of a law from the conjunction of itself and an unrelated law, or the deduction of a cause from one of its effects plus a law linking the two, such as the deduction of a galaxy's speed of recession from the red shift of its characteristic spectrum (cf. Psillos 2002: ch. 8). The red shift may entail the recession, but it is the recession that explains the red shift, not conversely. Causes are not the only things that are explanatory, but what makes them explanatory is not that they entail their effects. Indeed one of the signal strengths of the causal model of explanation is that it avoids so many of the weaknesses of the deductive-nomological model. It is however a weaknesses of the causal model that not all explanations fall within its purview.

3. ARE ALL CAUSES EXPLANATORY?

Each effect has many causes and not all of them explain it. When a student turns up to his tutorial without an essay written, the teacher may accept as at least a potential explanation the story about the computer crashing, but not 'Well, you know about the Big Bang ...'. The Big Bang is part of the causal history of every other event, but does not explain most of them. What then is the difference between explanatory and unexplanatory causes? One might look to the causes themselves. For example, a distinction is sometimes made between causes that are changes and causes that are standing conditions (Mill 1904:3. 5. 3): perhaps, the Big Bang notwithstanding, only changes are generally explanatory. Thus we might explain why the match lit by saying that it was because it was struck, not because there was oxygen present. But standing conditions are sometimes explanatory: we might for example explain why a match lit by saying that it was dry. Indeed the presence of oxygen may explain a fire, for example if the fire takes place in a laboratory environment that was designed to be oxygen-free (Hart and Honoré 1985:35).

One of the reasons we cannot distinguish between explanatory and unexplanatory causes in this way is because the distinction between changes and standing conditions is intrinsic to the cause, whereas the distinction between explanatory and unexplanatory causes is relative to the effect to be explained. The very same cause may explain one effect but not another. For example, the short circuit might explain the fire, but not why the insurance company refused to pay out. So we are better off looking for a demarcation that focuses on the relation between cause and effect. For example, overdetermined causes are often for that reason often not very explanatory. Thus if house is destroyed in an avalanche, mentioning the avalanche explains this better than mentioning only the particular rocks that happen to have hit the house, since if

those rocks hadn't hit the house, others would have. But many causes that are not overdetermined are still not explanatory, so this cannot be the whole story.

We can make some progress on the distinction between explanatory and un-explanatory causes by noting that what counts as an explanatory cause depends not only on the effect, but also on our interests. One natural way to account for the way interests help us to select explanations from among causes is to reveal additional structure in the why-question about the phenomenon to be explained, structure that varies with interest and that points to particular causes. The interest relativity of explanation can be accounted for in part with a contrastive analysis of what is explained. What is explained is not simply 'Why this?', but 'Why this *rather than* that?' (van Fraassen 1980: 126–9; Garfinkel 1981: 28–41; Lipton 2004: 30–54). A contrastive phenomenon consists of a fact and a foil, and the same fact may have several different foils. We may not explain why the leaves turn yellow in November *simpliciter*, but only for example why they turn yellow in November rather than in January, or why they turn yellow in November rather than turn blue.

Why-questions are often posed explicitly in contrastive form and it is not difficult to come up with examples where different people select different foils, requiring different explanations. Jones's untreated syphilis explains why he rather than Smith (who did not have syphilis) contracted paresis, but not why he rather than Doe (who had syphilis but had it treated) contracted paresis. An explanation of why I went to see *Jumpers* rather than *Candide* will probably not explain why I went to see *Jumpers* rather than staying at home, and an explanation of why Able rather than Baker got the philosophy job may not explain why Able rather that Charles got the job. Since the causes that explain a fact relative to one foil will not generally explain it relative to another, the contrastive question provides a restriction on explanatory causes that goes beyond the identity of the effect.

Although the role of contrasts in why-questions will not account for all the factors that distinguish explanatory from unexplanatory causes, it goes a considerable way. But how does it work: how does the choice of foil select an explanatory cause? There are various accounts available. Some focus on the ways in which an explanatory cause may probabilistically favour the fact over the foil (van Fraassen 1980: 146–51; Hitchcock 1999: 597–608). Others appeal to counterfactuals, requiring of an explanatory cause for example that it would not have been a cause of the foil, had the foil occurred (Lewis 1986: 229–30). A third kind of account is inspired by a classic principle of causal inference, John Stuart Mill's method of difference, his version of the controlled experiment (Mill 1904: 3. 8. 2). Mill's method rests on the principle that a cause must lie among the antecedent differences between a case where the effect occurs and an otherwise similar case where it does not. The difference in effect points back to a difference that locates a cause. Thus we might infer that contracting syphilis is a cause of paresis, since it is one of the ways Smith and Jones differed. The cause that the method of difference isolates depends on which control we use. If, instead of Smith, we used Doe, we would be led to say not that a cause of paresis is syphilis, but that it is the failure to treat it.

The method of difference concerns the discovery of causes rather than the explanation of effects, but the similarity to contrastive explanation is striking (Garfinkel 1981: 40). So there may be an analogous difference condition on contrastive explanation, according to which to explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the case of not-Q, where a

corresponding event is something that would bear the same relation to *Q* as the cause of *P* bears to *P* (Lipton 2004: 42–54). On this view, contrastive questions select as explanatory an actual causal difference between *P* and not-*Q*, consisting of both a presence and an absence. If only Jones had syphilis, that explains why he rather than Smith has paresis, since having syphilis is a condition whose presence was a cause of Jones's paresis and a condition that does not appear in Smith's medical history. The fact that *Jumpers* is a contemporary play and *Candide* is not caused me both to go to one and to avoid the other. Writing the best essay explains why Kate rather than Frank won the prize, since that is a causal difference between the two of them. So it appears that the reason some causes are not explanatory is that so many of our why-questions are contrastive, and for these only causes that mark a difference between fact and foil will provide good answers.

4. WHY DO CAUSES EXPLAIN?

An account of contrastive explanation can itself be seen as an answer to a philosophical contrastive question: why do some causes explain rather than others? But a good answer to this question may not explain why any causes explain, since it may simply presuppose that some causes do. To explain why any causes explain, we need to address different questions, which may also be contrastive, such as: why do causes rather than effects explain? For while some causes explain their effects, effects do not explain their causes. The recession of the galaxy explains why its light is red shifted, but the red shift does not explain why the galaxy is receding, even though the red shift may provide essential evidence of the recession, and either can be deduced from the other with the help of the Doppler law.

Do effects really never explain? Some good explanations appear at least superficially to be 'effectal'. Thus biologists appear to explain the presence of a trait in terms of its function, which is one of its effects. Thus we may explain the coloration of the wings of a moth in terms of its function of providing camouflage. Camouflage explains coloration, but camouflage is an effect of the coloration. It has, however, been argued that this appearance of effectal explanation is misleading, because functional explanations are actually causal. According to this 'selected effects' view of functional explanation, because of the natural selection mechanism, what explains current coloration is *past* camouflage. This caused the current coloration, because of the enhanced fitness that previous moths with such coloration enjoyed. So citing camouflage is to cite a cause after all (Wright 1976; Allen, Bekoff, and Lauder 1998). Perhaps there are other more plausible examples of legitimate explanation by effect, such as explanations by appeal to least-action principles in physics, but the explanatory asymmetry between cause and effect is very pronounced even if not quite universal, and an account of why this is so may help to show what makes causes explanatory.

The question as to why causes rather than effects explain is difficult to answer. It is difficult to avoid circular explanations, along the lines of, 'causes explain because they, unlike effects, have the power to confer understanding'. Moreover, there is a clear sense in which finding out about a thing's effects does increase our understanding of that thing. Indeed it may be that *P*'s effects typically tell one more about *P* than do its causes. For effects often give information

about *P*'s properties in a way that causes do not. This is so because physical properties are at least often dispositional, and dispositions are characterized by their effects and not by their causes. Thus to say that arsenic is poisonous is to say roughly that if you eat it you will die. The effects not only lead us back to the properties, but they are constitutive of at least some of them. In the conditional 'If you eat it, then you will die' there is both a cause and an effect, but they bear an asymmetrical relation to the corresponding property of being poisonous. Causing death is constitutive of the property of being poisonous, but eating arsenic, though a cause of death, is not constitutive of being poisonous. Nor do the causes of the arsenic or of its presence in a particular place appear to be constitutive of arsenic's properties. Yet the explanatory asymmetry between cause and effect still appears genuine: causes explain effects; effects do not explain causes.

A natural thought is that what is special about the causes of *P* is that they, unlike *P*'s effects, create or bring about *P*. Can this be the key to the explanatory asymmetry between causes and effects? But this may be another circular explanation. Why do causes explain effects? Because causes bring about effects. The worry is that 'bring about' is just another expression for 'cause', so all that has really been said is that causes explain because they are causes. A response would be to insist on a strong reading of 'bring about', a reading that would rule out a Humean account of causation, which takes causation to be no more than a constant conjunction or pattern of events.

Humeans may not like this, but they have the option of an error theory of explanation, according to which we never really explain why things happen, though the source of the illusion can be given, much as Hume himself had an error theory of necessary connection, according to which objects in the world are only conjoined, never connected, but the source of our mistaken idea of connection can be given. For Hume held that even though we cannot properly conceive of any connection between cause and effect, we nevertheless do have an idea of necessary connection. Hume traces that idea to the expectation we form of a familiar effect upon seeing a familiar cause. We then proceed illegitimately to project that feeling onto the world, supposing that the external cause and effect are themselves connected, in spite of absurdity of supposing that what connects one billiard ball to another on impact is a feeling of expectation (1748: sect. 7). Applied to the notion of explanation, such an approach would allow the reality of causation as pattern, but would treat understanding as a kind of pervasive illusion, since it depends on a notion of causation that is metaphysically untenable. This would still be to allow that our notion of explanation and understanding, however misguided, depends on the idea of things being created, generated, or brought about by their causes.

Many would find such eliminativism about understanding unpalatable. But an appeal to the thought that explanation depends on powerful metaphysical 'glue' linking *E*'s cause to *E* as a way of explaining why causes rather than effects explain might also be problematic for two other reasons. First, as an account of causation strengthens the link between an event and its causes, it will do likewise for the connection between an event and its effects, so it is not clear that an appeal to a strong connection between cause and effect actually helps to account for the explanatory asymmetry. Secondly, many good explanations appeal to causes that may not be strongly connected to what they cause. This is illustrated by explanatory causes that are omissions. A good answer to the question of why Jane is eating her campfire meal with a stick is that she has no spoon, yet there seems no especially strong metaphysical glue between the

absence of the spoon and the use of the stick.

A somewhat more promising answer to the question of why causes rather than effects explain appeals to the idea that only causes can make the difference between the phenomenon occurring and not occurring. This is connected to the idea of control, since effects are controlled through causes that make a difference, causes without which the effect would not occur. The causes of a phenomenon may be handles that could in principle have been used to prevent the phenomenon occurring in a way that the phenomenon's effects could not. To be sure, control is not always a practical option. The galaxy's recession causes and explains its red shift even though we are in no position to change its motion; but the speed of recession is nevertheless a cause that made the difference between that amount of red shift and another. This may partially account for why causes rather than effects explain, since causes often make a difference in this sense while effects never do. Information about causes provides a special kind of intellectual handle on phenomena because the causes may provide a kind of physical handle on those phenomena (cf. Woodward 2003).

One attraction of this view is that it may account for our ambivalence about the explanatory use of certain causes. For not all causes do make a difference. The obvious situation where they do not is one of overdetermination. An ecological example is an environment with foxes and rabbits (Garfinkel 1981: 53–6). To the question as to why a rabbit was killed we may answer by giving the location of the guilty fox shortly before the deed, or we may cite the high fox population in the region. Both are causes, but the details of the guilty fox's behaviour do not explain well because, given the high fox population, had that fox not killed the rabbit, another fox probably would have. Had the fox population been substantially lower, by contrast, the rabbit probably would have survived. The cause that made the difference is the cause that explains. The idea that causes explain because they provide a kind of handle is thus closely related to the difference condition on contrastive explanation discussed above. So it may be that one reason that some causes explain while others do not is in the end the same as the reason those causes explain while effects do not. Neither effects nor undiscriminating causes make the sort of difference between the phenomenon occurring and not occurring that provides understanding.

5. EXPLANATION AND CAUSAL INFERENCE

As we have seen, the metaphysics of causation may illuminate explanation. In turn, explanation may illuminate the epistemology of causation. This is the idea behind Inference to the Best Explanation: explanatory considerations are a guide to causal inference (Lipton 2004). Causal inferences are non-demonstrative, which means that there will always be competing causal hypotheses compatible with the same data. The suggestion is that we decide which of the competing hypotheses the evidence best supports by determining how well the competitors would explain that evidence. Many inferences are naturally described in this way. Seeing the ball next to the broken vase, the parent infers that the children have been playing catch in the house, because this is the best explanation of what the parent observes. Darwin inferred the hypothesis of natural selection because, although it was not entailed by his diverse biological evidence, the causal hypothesis of natural selection would provide the best

explanation of that evidence. Astronomers infer that a galaxy is receding from the earth with a specified velocity, because the recession would be the best explanation of the observed red shift of the galaxy's characteristic spectrum. Detectives infer that it was Moriarty who committed the crime, because this hypothesis would best explain the fingerprints, blood stains, and other forensic evidence. Sherlock Holmes to the contrary, this is not a matter of deduction. The evidence will not entail that Moriarty is to blame, since it always remains possible that someone else was the perpetrator. Nevertheless, Holmes is right to make his inference, since Moriarty's guilt would provide a better explanation of the evidence than would anyone else's.

Inference to the Best Explanation can be seen as an extension of the idea of self-evidencing explanations where the phenomenon that is explained in turn provides an essential part of the reason for believing the explanation is correct. For example, the speed of recession would cause and explain the red shift, but the observed red shift may at the same time be an essential part of the reason astronomers have for believing that the galaxy is receding at that speed. As we have seen, self-evidencing explanations exhibit a curious circularity, but this circularity is benign. The recession is used to explain the red shift and the red shift is used to determine the recession, yet the recession hypothesis may be both explanatory and well supported. According to Inference to the Best Explanation, this is a common situation: hypotheses are supported by the very observations they are supposed to explain. Moreover, on this model, the observations support the hypothesis precisely because it would explain them.

Inference to the Best Explanation thus partially inverts an otherwise natural view of the relationship between causal inference and explanation. According to that natural view, the inference is prior to the explanation. First we must decide which hypotheses to accept; then, when called upon to explain some observation, we will draw from our pool of accepted hypotheses. According to Inference to the Best Explanation, by contrast, it is only by asking how well various hypotheses would explain the available evidence that we determine which hypotheses merit acceptance. In this sense, Inference to the Best Explanation has it that explanation is prior to inference.

Although it gives a natural account of many inferences in both science and ordinary life, the model needs further development. What, for example, should be meant by 'best'? This is sometimes taken to mean likeliest or most plausible, but Inference to the Likeliest Explanation would be a disappointingly uninformative model, since the main point of an account of inference is to say what leads one hypothesis to be judged likelier than another, to give the symptoms of likeliness. A more promising approach construes 'best' as 'loveliest'. On this view, we infer the hypothesis that would, if correct, provide the greatest causal understanding.

The model should thus be construed as 'Inference to the Loveliest Explanation'. Its central claim is that loveliness is a guide to likeliness, that the explanation that would, if correct, provide the most understanding, is the explanation that is judged likeliest to be correct. This at least is not a trivial claim, but it faces at least three challenges. The first is to identify the explanatory virtues, the features of explanations that contribute to the degree of understanding they provide. There are a number of plausible candidates for the these virtues, including scope, precision, mechanism, unification, and simplicity. Better explanations explain more types of

phenomena, explain them with greater precision, provide more information about underlying causal mechanisms, unify apparently disparate phenomena, or simplify our overall picture of the world. But analysing these and other explanatory virtues is not easy, and it also leaves the other two challenges. One of these is to show that these aspects of loveliness match judgements of likeliness, that the loveliest explanations tend also to be those that are judged likeliest to be correct. The remaining challenge is to show that, granting the match between loveliness and judgements of likeliness, the former is in fact our guide to the latter.

In addition to offering a description of an important aspect of causal inferences, Inference to the Best Explanation has been used to justify them, to show that those causal hypotheses judged likely to be correct really are so. For example, it has been argued that there is good reason to believe that the best scientific theories are true, since the truth of those theories is the best explanation of their wide-ranging predictive success. Indeed it has been claimed that the successes of our best scientific theories would be inexplicable unless they was at least approximately true (Putnam 1978: 18–22).

This argument has considerable plausibility; nevertheless, it faces serious objections. If scientific theories are themselves accepted on the basis of inferences to the best explanation, then to use an argument of the same form to show that those inferences lead to the truth may beg the question. Moreover, it is not clear that the truth of a theory really is the best explanation of its predictive success. For one thing, it seems no better an explanation than would be the truth of a competing theory that happens to share those particular predictions. For another, to explain why our current theories have so far been successful may not require an appeal to truth, if scientists have a policy of weeding out unsuccessful theories.

The explanation that the truth of a theory would provide for the truth of the predictions that the theory entails appears to be logical rather than causal. This may provide some answer to the circularity objection, since the first-order scientific inferences that this overarching logical inference is supposed to warrant are at least predominantly causal. But it may also raise the suspicion that the real source of the plausibility of the argument is the plausibility of inferring from the premiss that most false causal hypotheses would have yielded false predictions to the conclusion that most causal hypotheses that yield true predictions are themselves true. Perhaps the premiss of this argument is correct, but the argument is fallacious. Most losing lottery tickets get the first three digits of the winning number wrong, but most tickets that get the first three digits right are losers too. It remains to be shown why the predictive successes of a general causal hypothesis is any better reason to believe that hypothesis is true than getting the first few digits of a lottery ticket right is a reason to think that ticket is a winner.

FURTHER READING

Lewis (1986) is an influential presentation of the view that to explain a phenomenon is to give information about its causal history. Lipton (2004) provides an accessible discussion of a causal model of explanation and of the idea that explanatory considerations are a guide to causal inference. Psillos (2002) is a clear introduction to causation, explanation, and the relations between the two. Salmon (1998) is a collection of essays on the relationship between causation and explanation by one of the most influential twentieth-century figures in this field. Woodward (2003) is a recent and detailed account of the relationship between causation and

explanation, emphasizing the importance of manipulation and control.

References

ALLEN, C., BEKOFF, M., and LAUDER, G. (eds.) (1998). *Nature's Purposes* . Cambridge, Mass.: MIT.

FRIEDMAN, M. (1974). 'Explanation and Scientific Understanding', *Journal of Philosophy* 71: 1–19.

GARFINKEL, A. (1981). *Forms of Explanation*. New Haven: Yale University Press.

HART, H., and HONORÉ, A. (1985). *Causation in the Law.* 2nd edn. Oxford: Oxford University Press.

HEMPEL, C. (1965). *Aspects of Scientific Explanation*. New York: Free Press.

HITCHCOCK, C. (1999). 'Contrastive Explanation and the Demons of Determinism', *British Journal for the Philosophy of Science* 50: 585–612.

HUME, D. (1748). *An Enquiry Concerning Human Understanding,* ed. T. Beauchamp. Oxford: Oxford University Press.

KAHNEMAN, D., SLOVIC, P., and TVERSKY, A. (eds.) (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

LEWIS, D. (1986). 'Causal Explanation', *Philosophical Papers II*. New York: Oxford University Press, 214–40.

LIPTON, P. (2004). *Inference to the Best Explanation*. London: Routledge.

MILL, J. S. (1904). *A System of Logic* 8th edn. London: Longmans, Green.

PSILLOS, S. (2002). *Causation and Explanation*. Chesham: Acumen.

PUTNAM, H. (1978). *Meaning and the Moral Sciences*. London: Hutchinson.

RUBEN, D. (1990). *Explaining Explanation*. London: Routledge.

SALMON, WESLEY (1984). *Scientific Explanation and the Causal Structure of the World* . Princeton: Princeton University Press.

—— (1998). *Causality and Explanation*. Oxford: Oxford University Press.

SORABJI, R. (1980). *Necessity, Cause, and Blame: Perspectives on Aristotle's Theory* . Ithaca: Cornell University Press.

VAN FRAASSEN, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

WRIGHT, L. (1976). *Teleological Explanations*. Berkeley: University of California Press.

WOODWARD, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

CHAPTER 30
CAUSATION AND REDUCTION

PAUL HUMPHREYS

1. INTRODUCTION

When considering the relations between causation and reduction one must distinguish between, on the one hand, issues about how causation operates within and between systems that stand in various reductive relations to one another; and on the other hand, issues concerning whether causation itself is amenable to a reductive treatment. These two issues are intertwined and each must be treated with sympathy for the other. There are two basic types of reduction. Ontological reduction concerns reductive relations between the objects themselves whereas linguistic or conceptual reduction deals with reductive relations between our representations of those objects. For the great majority of the last century, both causation and reduction were treated linguistically or conceptually, but in recent years there has been a significant shift towards directly ontological treatments of each.

2. ELIMINATIVIST AND NON-ELIMINATIVIST REDUCTION

The basic idea behind reduction in many of these approaches is that if X can be reduced to Y, X is nothing but Y. Inevitably, there are dissenters from this view of reduction who, for example, claim that a property F has been reduced to other properties G if F has been completely explained in terms of G (see e.g. Wimsatt 2006) but we shall not deal with that approach here. Within the standard approaches 'nothing but' has at least two senses. In *eliminative reduction,* the reduced entity or concept A, which is often considered to be disreputable by those attempting the reduction, is shown to be dispensable and hence can be eliminated from our ontology or from our theoretical discourse. Historically, the majority of empiricists have held that causation itself requires an eliminativist treatment because causal relations, and relations of causal necessitation in particular, have been taken to be inaccessible to the kinds of reasonably direct observation championed by empiricists. They therefore need to be reduced to elements of reality that are more acceptable to empiricists. Here the shadow of Hume looms large. Non-eliminativists, such as Armstrong (1968) and Fales (1990),[1] hold that causal relations are primitive in some sense and that we can in certain cases have direct epistemic access to them.

Included in the eliminativist school were most philosophers working in the logical empiricist tradition, including influential philosophers of science such as Ernest Nagel and

Carl Hempel. Many contemporary physicalists also consider themselves eliminativists. Probably the most notorious eliminativist was Bertrand Russell, whose views are encapsulated in the famously barbed remark, 'the reason why physics has ceased to look for causes is that, in fact, there are no such things. The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm' (Russell 1918: 174). A less confrontational approach was the logical replacement tradition, dominant in the middle years of the last century, within which the sentence 'event a caused event b' is true if and only if there are law-of-nature statements $L_1, ..., L_n$ such that from those statements and a sentence describing the occurrence of a, a sentence describing the occurrence of b can be deduced. Causal necessitation in this approach was replaced by the logical necessitation possessed by the deductive relation. One such typical account can be found in Hempel (1965).

In *non-eliminativist reduction*, X is reduced to Y but is not eliminated, X being allowed to remain as a legitimate and practically useful, although ultimately redundant, element of our discourse or ontology. The point of a non-eliminativist reduction is often to heighten our understanding of the reduced item by providing a theory or an analysis of the item, with the analysis being either logical or structural. Many contemporary philosophers have adopted non-eliminativist approaches to causation on the grounds that causation is too central to our accounts of explanation, experiment, and action to suffer elimination, while at the same time they seek to lessen its mysterious nature by analysing it in terms of other, more acceptable, concepts such as statistical associations, actual or hypothetical interventions in the world, or lawlike regularities.

A third important type of reduction amends the reduced concept in the course of carrying out the reduction, thus providing a *revisionary reduction* of causation. The reducing theory may, for example, supply a more precisely defined domain for the target concept or offer superior criteria for identifying instances of the concept. In such cases the original concept may be either discarded or kept as a false but pragmatically useful tool. Thus our ordinary, everyday concepts of causation may loosely fit the world in some areas, but for scientific or philosophical purposes they may well need to be replaced by more refined accounts of a universal nature. It is important to distinguish between a revisionary reduction that rejects certain widely accepted examples of causation as in fact not being cases of genuine causation at all, a position that is possible but needs a considerable amount of ancillary justification, and a revisionary reduction that fails to fit common linguistic claims involving causal relations. The linguistic failure, assuming it does not mirror a failure to capture the causal facts but simply does not capture all the causal vernacular, is not a relevant constraint for contemporary ontologically oriented reductions. The expression 'His argument caused consternation in the crowd' by itself cuts no ice with the issue of whether reasons are causes and a revisionary physicalist reduction of causes can happily make that expression literally false and suffer no consequences. Most philosophical accounts of causation are revisionary to a greater or lesser degree; a revisionary reduction from a scientific perspective can be found in Pearl (2000).

Whether a treatment of causation is revisionary or non-revisionary, it is often forced to use in its analysis concepts that, although not obviously causal, may ultimately require an appeal to causation fully to account for them. Such concepts include natural necessity, counterfactual conditionals, and dispositions. This is the famous modal circle from which escape has proved

to be extremely difficult. In fact, some accounts of causation concede that no fully satisfactory account of causation is possible that eliminates references to causes altogether. For example, within Cartwright's account of probabilistic causality (Cartwright 1983: 26), in order for a relation to be causal, it has to be invariant under changes in other causal relations, thus referring to causes in the definiens as well as in the definiendum.

Some care must be taken in classifying an account as eliminativist or non-eliminativist since the term 'causes' can fail to refer in its original domain of application but can do so successfully in another. This is the situation with Hume's (1739) famous empiricist dismissal of natural necessities which carries out a revisionary reduction of causation. For Hume, humans are inclined to attribute causation to a sequence of events just in case those events are part of an observable regularity of similar events, the cause event temporally precedes the effect event, and the cause-and-effects events are spatially and temporally contiguous.[2] Each of these three conditions is free from causal content[3] and each has the added advantage of conforming to the constraints of an empiricist epistemology. The fourth condition suggested by Hume, that we have a psychological habit, acquired from observing the regularity, of passing from the idea of an event similar to the cause to the idea of an event similar to the effect, eliminates causal necessity from the world but allows us to make causal attributions on a psychological basis.

3. DOMAINS OF REDUCTION

Discussions of reduction are often based on the assumption that there is a hierarchy of subject matters, each with its own level of ontology, with physics at the foundation of the hierarchy, chemistry above physics, then molecular biology and biochemistry, cellular biology, and so on, up through neuroscience to the various social sciences. This assumption of a hierarchy of levels can prejudice the treatment of some reductive issues and the neutral language of reducing one domain to another, suggested in Kim (1999), is to be preferred.

In addition to *domains*, attempts at reduction can also take place between *theories* such as thermodynamics and statistical mechanics; between *terms* in theories, such as 'water' and 'H_2O'; between *objects* such as social groups and their members; between *laws* of different sciences; and between *properties* such as brittleness and a particular molecular structure.[4] These differences are important in the case of causation because different theories of causation take different kinds of entities as their primary focus—laws, properties, and event names being well-known candidates. For example, if it turns out that there are no laws in a given domain, then if laws are necessary to ground causal relations, no reduction of causation from another domain into that domain will be possible, whereas if singular causal relations are taken as the primitives of one's account, a reduction is not precluded simply by the absence of laws.

4. RELATIONS BETWEEN DOMAINS

With this apparatus in place, we can address in greater detail the relations between causation and reduction. Candidates that have been proposed for the relations that hold between causal elements in different domains include: a relation of identity; an explicit definition; an implicit definition via a theory; a contingent statement of a lawlike connection, as in Nagel-reduction; a relation of natural or metaphysical necessitation, as in supervenience; an explanatory relation, as in Kim-reduction; a relation of emergence; a realization relation; a relation of constitution; and even causation itself. I shall not discuss in any detail the first three of these since the relations involved are familiar and they do not raise any special issues for causation and reduction.

5. DOMAIN-SPECIFIC CAUSATION AND PHYSICALISM

An issue of some interest is whether there are types of causation that are specific to particular subject matters or whether all causation is amenable to a single comprehensive analysis. The most discussed examples of potentially distinctive types of causation are the case of mental causation (e.g. Davidson 1980), which, if it exists, may have special features such as intentionality, and the associated concept of agency causation (e.g. Frankfurt 1988) which may be different in kind from event causation. One should also keep open the possibility of there being special modes of causal action in some of the social sciences. Examples such as these raise the question of whether there could be reductive relations between domain-specific types of causation and the effects that the existence of such special forms of causation would have on claims about the causal closure of the physical realm.

Many accounts of causation based on conceptual analysis, notably counterfactual analyses (see e.g. Lewis 1973; 1986*a*; Woodward 2003) and regularity analyses, aspire to subject-matter-independent status. They can thus aim at universality in virtue of not committing the account to a special underlying ontology, except for a generic use of events or propositions. In particular, nothing in those accounts precludes an application to non-physical entities such as mental events. But even some of these very general accounts are not totally universal. If causation can hold between abstract Platonic entities (as is suggested in Cargile 2003, for example), the necessary existence of such entities would preclude the application of a counterfactual analysis of causation unless a satisfactory semantics for counter-factuals that is applicable to impossible worlds can be constructed.

In contrast, of growing importance in recent years has been a particular sort of domain-specific claim about causation and other matters in the form of *physicalism*, the position that the only genuine existents are those described by the true theories of the physical sciences.[5] An influential branch of physicalism contains the various programmes known as *Humean supervenience*, discussed in sect. 7 below. Although all analyses of causation, with the possible exception of divine interventionalism, allow causation to occur between physical events, some, such as those of Fair (1979), Wesley Salmon (1998), and Dowe (2000) which ground causal relations in conserved quantities from physics, implicitly restrict causation to physical causation. In so doing, they make such a thing as mental to physical causation

improbable in our world because it would require the transfer of conserved physical quantities such as energy from the mental to the physical in a way that has never been empirically detected. Also, although it is possible that there might be some specifically mental or social quantities that are conserved and so would allow this type of causation to hold between mental or social events, the accounts just mentioned would only accidentally capture such things because the cited authors clearly intend the causally relevant properties to be physical in form. The adoption of such a physicalist approach has far-reaching consequences for one's attitude towards causation in the social sciences, but if one is both sympathetic to eliminativism in the philosophy of mind and denies the existence of specifically social facts and others of that ilk, the position may well resonate. Indeed, this consequence has been explicitly embraced by Merrilee Salmon (2003).

If a physicalist reduction of all forms of causation is possible, then the conceptually based, subject matter independent approach and the physicalist approach to causation will both be universal but they will in most versions differ in that the physicalist theories aspire only to contingent truth, whereas the more general approaches aspire to necessary truth—these general accounts of causation are supposed to be correct about causation independently of the way the world might be. The contingency of the physicalist accounts must, however, be distinguished from the modal status of relations between causal connections in different domains. That is, social causation might well depend conceptually upon physical causation even though the theory of physical causation was itself only contingently true.

6. Contingent Relations

The literature on reduction was for many years dominated by Ernest Nagel's (1961) account of reductive relations between theories. Although his account is framed in terms of theories, Nagel's interest in reduction was in fact motivated by his desire to reduce relations between properties that were peculiar to one domain in favour of dependency relations between properties that were not unique to that domain—as he put it 'certain relations of dependence between one set of distinctive traits of a given subject matter are allegedly explained by, and in some sense "reduced" to, assumptions concerning more inclusive relations of dependence between traits or processes not distinctive of (or unique to) that subject matter' (Nagel 1974: 96). It is enlightening to see how causal relations fare under Nagel-reduction. We need be concerned here only with what Nagel called inhomogeneous reductions for causation—that is, reductions in which the vocabulary special to the reduced theory does not occur in the reducing theory—because if one is an eliminativist, one will not have causal vocabulary in the reducing domain or, if there are different types of causal relation in different domains, there will initially be different causal vocabularies such as 'social causation' and 'physical causation' in the different domains.

In the case of inhomogeneous reductions, Nagel suggested that the reduction must be carried out by constructing bridge laws. For Nagel, the bridge laws are contingent empirical hypotheses and not definitions. As Nagel notes, the extensional equivalence between terms that occurs with some bridge laws allows the existence both of the bridge laws and of the

things picked out by both relata, even though the meaning of a term such as 'viscous' will be different from the meaning of the extensionally equivalent term 'frictional forces between layers of molecules'. This way of framing the issue is important for causal reduction because typographically identical occurrences of the term 'causes' in different domains may turn out to be homonyms rather than synonyms. One option is to maintain that the meaning of the term 'causes' should be the same in all domains, regardless of whether it is prefixed by a modifier such as 'social'. That is, although the mechanisms by means of which social causes, if they exist, bring about their effects may be different from the ways in which physical causes bring about their effects, there should be a single sense of causation involved across both domains if the same term is to be used in each, just as there is within the entire realm of physical causation. The other option is to allow that there are distinctively different types of relation in different domains, each of which bears the title 'causal' only by loose association and has at least one feature that is peculiar to that domain. If this is so, then if there is a distinctive type of social causation, for example, the relation between tokens of that type of causation and corresponding tokens of physical causation would not itself be purely physical because one element of the relation between the tokens, the social causal relation, would be non-physical.

7. SUPERVENIENCE AND CAUSATION

The greatest challenge to Nagel-reduction comes from the phenomenon of multiple realizability, which is discussed in greater detail in sect. 8. The *locus classicus* of this challenge can be found in the influential argument against the type–type identity of properties developed in Fodor (1974), the essential point being that the multiple realizability of a property F by properties G_1, ..., G_n allows us to reduce F only to the disjunction G_1 v ... v G_n rather than to a single natural property of the reducing domain. Because Nagel-reduction involves a reduction of one theory to another, there will not in general be a scientifically acceptable theory containing heterogeneous disjunctive predicates of that kind. For present purposes, a key issue is that Fodor's argument does allow the identification of instances of causal properties and so does not block a reduction of token causal relations. His argument is thus consistent with a single case account of causation that does not require scientific laws for causal attribution.

One response to the multiple realizability argument has been the development of an influential but controversial set of supervenience accounts of the relations between various domains. A rough characterization of supervenience is:

> One family of properties F is supervenient upon another family of properties G if (i) two things alike for all members of G must be alike with respect to all members of F but (ii) there is no relationship of definability or entailment between the two families.

Of the variety of definitions of supervenience available, this one will serve:

A set of properties *M* strongly supervenes on a family *N* of properties iff, necessarily, for each property *F* in *M*, if *F*(*x*) then there is a property *G* in N such that *G*(*x*) and necessarily if any *y* has *G* it has *F*.

A particular variety of supervenience-based physicalism, *Humean supervenience*, is a key element of many modern formulations of physicalism. Roughly, it asserts that the world consists in a spatio-temporal distribution of localized physical particulars and everything else, including laws of nature and causal relations, supervenes on that. Here, for example, is Tooley's definition of strong reductionism with respect to causal relations:

Any two worlds that agree with respect to all of the non-causal properties of, and relations between, particulars, must also agree with respect to all of the causal relations between states of affairs (causal laws). Causal relations (causal laws) are, in short, logically supervenient upon non-causal properties and relations. (Tooley 2003: 388)

Humean supervenience can be held as a necessarily true or as a contingently true thesis. Arguments due to Tooley (1987: 47–8) and Carroll (1994: ch. 3) have shown that Humean supervenience is at best contingently true of laws (and hence, in law-based accounts of causation, of causation as well). We shall thus restrict our attention to Humean supervenience as a contingent claim about our world. In this version, we have David Lewis's famous statement:

[Humean Supervenience] is the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another. (But it is not part of the thesis that these local matters are mental.) We have geometry: a system of eternal relations of spatiotemporal distance between points. Maybe points of spacetime itself, maybe point-sized bits of matter or aether or fields, maybe both. And at those points we have local qualities: perfectly natural intrinsic properties which need nothing bigger than a point at which to be instantiated. For short: we have an arrangement of qualities, and that is all. There is no difference without difference in the arrangement of qualities. All else supervenes on that. (Lewis 1986*b*: pp. ix–x)

Although supervenience became popular as a way to achieve some of the physicalists' goals, there is considerable, although often implicit, disagreement amongst advocates of this approach over the degree to which supervenience results in a non-reductive relation. One can find a number of remarks in the literature asserting that if *X* supervenes on *Y* then *X* is nothing but *Y*. Nevertheless, supervenience is more generally conceived of as a non-reductive relation, where the term 'non-reductive physicalism' is used to allow an agnostic position on the issue of reducibility or irreducibility.

8. FUNCTIONALISM AND REALIZABILITY

Multiple realizability occurs when properties are functionally characterized in terms of their causal roles. For example, the property of serving as money can be causally realized by banknotes, by electronic transfers, by company store tokens, by gold, by beads, and in many other ways.[6] A sizeable literature has now developed suggesting that multiple realizability is much less common than has been supposed and that there are ways in which multiple realizability can be made compatible with reduction. Consider the case of doorstops. These are functionally characterized and can be realized by bricks, by cast-iron pigs, by wooden wedges, by bean bags, and so on. Yet the relevant causal factor has been isolated by physics as a very general property, the property of providing a counteracting force of magnitude at least as great as that exerted by the door. The simplicity of the doorstop example reveals something of importance for it shows that the existence of multiple realizability does not inevitably undermine the possibility of reductionism. Although there are many specific forces such as those resulting from friction, torsion, leverage, and so on, each of which itself realizes the more general property just noted, the identification of the general property involved constitutes an important reductive step because there does exist a physical property, that of a counteracting force of magnitude at least M, that is present in all cases where the functional property is present. Furthermore, it is a contingent truth that a force of magnitude M will physically counteract a force of the same magnitude acting in the opposite direction—it is logically possible for there to be worlds in which force laws are spatially asymmetric.

Here is a more detailed case. It was once common to characterize acids functionally, as substances that produced carbon dioxide when added to sodium or calcium carbonate, or that turned red phenolphthalein colourless. With that characterization, acidity is realized by acetic acid, by sulphuric acid, by hydrochloric acid, and so on, a set of substances with a fairly heterogeneous collection of properties. The contemporary Lewis concept of an acid, however, defines it as a species that can accept an electron pair. Now we have a common property that is realized in acetic acid, in sulphuric acid, and in hydrochloric acid. In virtue of this common feature, we have a reduction of a wide variety of functionally characterized substances to a common property, and moreover a common property that is causally responsible for the effects associated with acids. The moral is that the inference from multiple realizability to irreducibility can only be made on a case by case basis, not generally.

9. REDUCTION AND DOWNWARD CAUSATION

Two influential arguments that make it difficult to maintain the possibility of causal influence in domains outside the realm of the physical are the exclusion argument and the downward causation argument. The exclusion argument has three premises:

(1) If an event a is causally sufficient for an event b, then any event c distinct from a is causally irrelevant to b.

(2) For every physical event *b*, there is some physical event *a* that is causally sufficient for *b*.

(3) If an event *c* supervenes on a physical event *a*, then *c* is distinct from *a*.

The conclusion is then:

For every physical event *b*, no supervening event *c* is causally relevant to *b*.

The exclusion argument can be applied to any kind of event that supervenes upon physical events and it shows that there is no obvious causal role in the world for supervening events. As Samuel Alexander wrote, in a passage curiously evocative of Russell's: 'Epiphenomenalism supposes something to exist in nature which has nothing to do, no purpose to serve, a species of noblesse which depends on the work of its inferiors, but is kept for show and might as well, and undoubtedly would in time, be abolished' (Alexander 1920: 8, quoted in Kim 1996: 129).

The downward causation argument has a different, albeit closely related, orientation. To present it, we must temporarily switch back to the representation of properties as arranged in a hierarchy of levels. The argument is designed to show that, given this levels picture, any instance of causation at a level above the most basic must be accompanied by downward causation, that is, a causal relation from the higher to the lower level. Call the higher-level property instance *F*, the lower-level property instance *G*, and suppose that non-reductive physicalism is committed to the view that in order for any such *F* to occur, it must be realized by, that is, simultaneously determined by, some *G*. Now allow *F* to cause some later *F**. This *F** will have its own realizer *G** and *G** determines *F**. Given this situation, unless *F*'s causation is to be redundant, *F* must bring about *F** by bringing about *G**, which realizes *F**. And *F*'s bringing about *G** is precisely an instance of downward causation. The argument is general, and so higher-level causation, within layered systems for which realization holds, is accompanied by downward causation. If one then combines the downward causation argument with the exclusion argument, there is serious trouble for the higher level in that the downward causation argument requires *F* to be causally relevant to *G**, whereas the exclusion argument concludes that it cannot be. Now employ what has come to be known as *Alexander's Dictum,* 'To be real is to have causal powers' and the net conclusion is that the higher-level events are not genuine existents. So these causal considerations lead to a rejection of the irreductive aspect of non-reductive physicalism. It would seem that only reductive physicalism is tenable.

Alexander's Dictum, reformulated in what I take to be an equivalent form, asserts that an entity is real if and only if has at least one causal power. Even as a sufficient condition for something to be real, this version of Alexander's Dictum is not entirely unobjectionable, since nominalists could maintain that only individuals are real, individuals almost always have multiple causal powers, and the revised dictum allows single causal properties as real. As a necessary condition, it is unavoidably controversial. Those who believe in the reality of abstract objects and who hold that one characteristic of abstract objects is that they are not causally efficacious will reject Alexander's Dictum. Also, as Eugene Mills (2003) has noted,

the dictum makes epiphenomenalism impossible, in that the effects of causes seem to be legitimate existents, but if those effects, as epiphenomena, do not themselves cause anything, by the dictum they do not exist. In the exclusion argument, Alexander's Dictum is used as a necessary condition for existence, but it seems plausible that we can avoid appeal to it by considering in any given case whether it is part of our conception of the putatively existing object that it exerts causal influence (see Elder 2003). That surely is true of most people's conception of mental properties and in general it would be enough to label any entity that was the victim of a successful exclusion argument to be epiphenomenal. Whether one then wanted to deny its existence would be a matter for specific consideration.

10. REDUCTION AND EMERGENCE

The primary rival to reduction is emergence, and reduction and emergence are generally considered to be incompatible. (For an opposing view, see Wimsatt 2006.) Interestingly, there are very few accounts of emergence that are explicitly causal in form. This is because within the traditional unpredictability accounts of emergence (e.g. Broad 1925) and the contemporary computationally based complexity-based approaches to emergence (e.g. Bedau 2002) causation plays no role whatsoever, and conceptual emergence is, as its name indicates, concerned with the emergence of novel representational frameworks. This leaves ontologically based accounts of emergence and within those, it seems to be widely if implicitly assumed that ordinary causal relations are too well understood to produce anything like emergent features. In consequence, attention tends to be focused on the realm of the mental, which as we have seen produces its own special problems, or on quantum mechanical and other physical phenomena, within which formal rather than causal considerations are primary. Exceptions to this scarcity are a specifically causal version of dynamic emergence found in O'Connor and Wong (2005) and a thinly sketched form of causal emergence defended in Searle (1992). Despite this generally dismissive attitude towards traditional causation in the emergence literature, it is worth raising the following question (see Humphreys 1997*b*): why cannot traditional 'horizontal' causation give rise to emergent features that are present within the same domain of properties as gave rise to them? Specifically causal accounts of emergence should surely leave open this possibility and if such things exist, the downward causation argument is no longer a problem.

One route through which causation does enter treatments of emergence is through the standard requirement that an emergent property be novel, because one way for a property to be novel is for it to possess novel causal powers. The novelty of the causal powers follows from the novelty of the property if we accept Shoemaker's (1980) identity criterion for properties, that a property is individuated by the powers it contributes to individuals. Yet once novel causal powers are entertained, the exclusion and downward causation arguments must be effectively handled. One way to do this is explored in Humphreys (1997*a*).

An important alternative to Nagel-reduction has been developed by Jaegwon Kim specifically to address the issue of emergence (Kim 1997; 1999). As Kim puts it, 'Functionalization of a property is both necessary and sufficient for reduction (sufficient at

least as a first conceptual step, the rest being scientific research) … it explains why reducible properties are predictable and explainable' (Kim 1997: 13). In order to Kim-reduce a property *F* to a set of properties *G* we must:

1. Functionalize *F*, i.e. construe *F* as a property defined by its causal/nomic relations to other properties in the reduction base *G* (either directly or through further reductions)

2. Find realizers of *F* in *G*.

3. Find a *G*-level theory that explains how realizers of *F* perform the causal task constitutive of *F*.

One of the characteristic features of Kim's approach to reduction is that, unlike Nagel's account, it does not require that bridge laws be found between the two levels and it is immune to problems arising from multiple realizability. The bridge laws are replaced by conceptual connections between *F* and the properties in *G* that are created during the functionalization process. That is, the higher-level property is conceived of as being the intermediate link between the causal input and the causal output at the lower level.

An important conclusion of Kim's arguments is that almost all properties can be functionalized and hence very few, perhaps only qualia and consciousness, are candidates for emergent properties. For arguments against Kim's position see Anthony (1999) and Shoemaker (2002).

11. SUMMARY

It is probably safe to say that a schism has developed between metaphysically oriented accounts of causation within which a reduction of causal relations to non-causal features is believed to have been achieved, and more scientifically oriented accounts within which those reductions are either viewed with suspicion or rejected in favour of approaches appealing to other causal concepts. Most of the main lines of research within this second area seem to be converging, albeit in different ways, on non-reductive, revisionary, non-eliminative accounts of causation. Significant progress has been made in understanding what relations are possible between causes operating in different domains with perhaps the most interesting new directions involving how emergent entities can play a causal role in the world. The great unanswered question is in what sense, if any, the material realm is causally closed.

FURTHER READING

Bedau and Humphreys (2008) contains a variety of approaches to emergence and reduction including articles by Nagel, Fodor, Hempel, Kim, Bedau, Humphreys, Wimsatt, and Searle mentioned in this chapter. It also has detailed overviews of scientific and philosophical

approaches to the topic. Carroll (1994) offers an original and lucid position on nomological reduction and is a good entry point into the large literature on the topic. Gillett and Loewer (2001) contains articles by many prominent contemporary physicalists arguing for or against reductionist positions, including defenders of Humean supervenience positions. Kim (1993) is a collection of articles by the author that provides both an in-depth account of supervenience and a variety of problems about causation with which non-reductive physicalism is faced. Woodward (2003) is a comprehensive recent treatment in the area, arguing for an interventionist counterfactual position on causation. It is lengthy, but a useful reference source.

REFERENCES

ALEXANDER, SAMUEL (1920). *Space, Time, and Deity*. London: Macmillan, ii.

ANTHONY, LOUISE (1999). 'Making Room for the Mental', *Philosophical Studies* 95: 37–44.

ARMSTRONG, DAVID (1968). *A Materialist Theory of the Mind.* London: Routledge & Kegan Paul.

BEDAU, MARK (2002). 'Downward Causation and Autonomy in Weak Emergence', *Principia Revista Inernacional de Epistemologica* 6: 5–50.

—— and HUMPHREYS, PAUL (eds.) (2008). *Emergence: Contemporary Readings in Science and Philosophy*. Cambridge: Mass.: MIT.

BROAD, CHARLES (1925). *The Mind and Its Place in Nature*. London: Routledge & Kegan Paul.

CARGILE, JAMES (2003). 'On "Alexander's" Dictum', *Topoi* 22: 143–9.

CARROLL, JOHN (1994). *Laws of Nature*. Cambridge: Cambridge University Press.

CARTWRIGHT, NANCY (1983). 'Causal Laws and Effective Strategies', *How the Laws of Physics Lie*. Oxford: Oxford University Press, 21–43.

DAVIDSON, DONALD (1980). 'Mental Events', *Essays on Actions and Events*. Oxford: Clarendon, 207–25.

DOWE, PHIL (2000). *Physical Causation*. Cambridge: Cambridge University Press.

ELDER, CRAWFORD (2003). 'Alexander's Dictum and the Reality of Familiar Objects', *Topoi* 22: 163–71.

FAIR, DAVID (1979). 'Causation and the Flow of Energy', *Erkenntnis* 14: 219–50.

FALES, EVAN (1990). *Causation and Universals*. London: Routledge.

FODOR, JERRY (1974). 'Special Sciences: Or, the Disunity of Science as a Working Hypothesis', *Synthese* 28: 97–115.

FRANKFURT, HARRY (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

GILLETT, CARL, and LOEWER, BARRY (eds.) (2001). *Physicalism and Its Discontents*. Cambridge: Cambridge University Press.

HEMPEL, CARL (1965). 'Aspects of Scientific Explanation', *Aspects of Scientific Explanation and Other Essays*. New York: Free Press.

HUME, DAVID (1739). *A Treatise of Human Nature*. (Various modern editions, see e.g. ed.

David Fate Norton. Oxford: Oxford University Press, 2001.)

HUMPHREYS, PAUL (1997*a*). 'How Properties Emerge', *Philosophy of Science* 64: 1–17.

—— (1997*b*). 'Emergence, Not Supervenience', *Philosophy of Science* 64: S337–S345.

KIM, JAEGWON (1993). *Supervenience and Mind*. Cambridge: Cambridge University Press.

—— (1996). *Philosophy of Mind*. Boulder: Westview.

—— (1997). 'Explanation, Prediction, and Reduction in Emergentism', *Intellectica* 2: 45–57.

—— (1999). 'Making Sense of Emergence', *Philosophical Studies* 95: 3–36.

LEWIS, DAVID (1973). 'Causation' *Journal of Philosophy* 70, pp. 556-567.

—— (1986*a*). 'Postscripts to "Causation"', in Lewis (1986*b*: 172–213).

—— (1986*b*). *Philosophical Papers II*. Oxford: Oxford University Press.

MILLS, EUGENE (2003). 'An Epistemic *Reductio* of Causal Reductionism', *Topoi* 22: 151–61.

NAGEL, ERNEST (1961). *The Structure of Science*. New York: Harcourt, Brace & World.

—— (1974). 'Issues in the Logic of Reductive Explanations', *Teleology Revisited and Other Essays in the Philosophy and History of Science*. New York: Columbia University Press, 95–113.

O'CONNOR, TIMOTHY, and WONG, HONG YU (2005). 'The Metaphysics of Emergence', *Noûs* 39: 658–78.

PEARL, JUDEA (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

RUSSELL, BERTRAND (1918). 'Mysticism and Logic', *Mysticism and Logic and Other Essays*. New York: Longmans Green.

SALMON, WESLEY (1998). *Causality and Explanation*. Oxford: Oxford University Press.

SALMON, MERRILEE (2003). 'Causal Explanations of Behavior', *Philosophy of Science* 70: 720–38.

SEARLE, JOHN (1992). 'Reductionism and the Irreducibility of Consciousness', *The Rediscovery of the Mind*. Cambridge: MIT, 111–26.

SHOEMAKER, SIDNEY (1980). 'Causality and Properties', in Peter van Inwagen (ed.), *Time and Cause*. Dordrecht: D. Reidel, 109–35

—— (2002). 'Kim on Emergence', *Philosophical Studies* 108: 53–63.

TOOLEY, MICHAEL (1987). *Causation*. Oxford: Clarendon.

—— (2003). 'Causation and Supervenience', in M. Loux and D. Zimmerman (eds.),*Oxford Handbook of Metaphysics*. Oxford: Oxford University Press.

WIMSATT, WILLIAM C. (2006). 'Emergence as Non-Aggregativity and the Biases of Reductionisms', in William C. Wimsatt, *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, Mass.: Harvard University Press.

WOODWARD, JAMES (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

PART VII
CAUSATION IN OTHER DISCIPLINES

CHAPTER 31
CAUSATION IN CLASSICAL MECHANICS

MARC LANGE

1. Russell's Causal Eliminativism

Russell spoke for many twentieth-century philosophers (and late nineteenth-century physicists, such as Mach) in suggesting that classical physics had discovered the universe to be an acausal place: 'The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm' (1912–13/1953: 387). In Russell's austere view, the laws of classical physics are merely relations holding among various physical quantities instantiated at various times. The laws do not portray certain of these property instantiations as causes of certain others. For example, Newton's second law of motion (F = ma) relates a body's acceleration (a) at a given moment to the body's mass (m) and the net force (F) on the body at that instant. The equation does not privilege one of these quantities as an effect of the other two. To regard acceleration, for example, as caused by force and mass would be to regard force as like muscular push or pull. The bare equations of physics are cleansed of any such anthropomorphic vestige and (to generalize Hertz's (1893: 21) remark regarding Maxwell's electromagnetic theory) a given physical theory is just its system of equations.

Russell might be interpreted as arguing that physics reveals there to be no causal relations since physics has no need to posit them (just as Laplace said to Napoleon that physics has no need to posit God). Of course, whether physics needs to posit causal relations depends upon what physics needs to do. Russell appears to presume that a physical theory needs merely to predict certain quantities from others. For that purpose, the bare equations suffice. However, it is doubtful that the bare equations are enough to fund scientific explanations. The force on a body and the body's mass apparently explain why the body undergoes a given acceleration, whereas the force and acceleration do not explain why the body possesses a certain mass. The notorious difficulties (reviewed by Salmon 1989) encountered by Hempel's (1965) D-N model of scientific explanation reveal how difficult it is to explicate explanatory priority without appealing to causal priority.

Arguably, physics also aims to tell us how to bring about the conditions we desire (see Cartwright 1983). To make a 1-kilogram body accelerate at 5 metres per second per second, we should shove it with 5 Newtons of force. But if instead we wanted to know how to put a 5-Newton force on the body, the advice 'Make the body accelerate at 5 m/s^2' would be unhelpful. That the body's acceleration happens to equal 5 m/s^2 might tell us that the net force that we have somehow already managed to apply equals 5N. But it would not suggest how to

arrange such a force in the first place.

In response, Russell might deny that physics aims to explain events or to tell us how to bring them about. Questions about how richly or austerely to interpret physics are often difficult and will plague us again. In any event, Russell apparently regards classical physics as saying more about certain equations than that they hold true. He regards classical physics as deeming them to be natural laws rather than accidents such as the fact that all gold cubes are smaller than a cubic mile (presuming this to be true) and the fact that all mammals contain DNA (an accident of evolutionary history). That causal relations fail to figure explicitly in the bare equations cannot be Russell's only reason for maintaining that classical physics reveals there to be no causal connections (else he would also have to maintain that classical physics reveals there to be no lawhood). Shortly we will see some other possible reasons.

2. CAUSAL LAWS AS DIFFERENTIAL EQUATIONS

Some philosophers have challenged Russell's eliminativism by contending that causal relations *can* be straightforwardly read off of the bare nomic equations. Frank (1932/1998: 143–6), for example, held that the fundamental laws of classical physics are differential equations—that is, equations giving one quantity's instantaneous rate of change at time t as a mathematical function of various standing conditions at t. We predict quantity q's value at a later time t_2 from its value at an earlier time t_1 and its instantaneous rate of change (dq/dt) at each moment in between:

$$q(t_2) = q(t_1) + \int_{t_1}^{t_2} [dq(t)/dt]\,dt.$$

Apparently, then, $q(t_2)$ is the cumulative effect of q's instantaneous rates of change during the intervening period coupled with the initial condition $q(t_1)$, and each of these instantaneous rates of change, in turn, is caused by the standing conditions identified by the relevant differential equation. Which quantities are causes of which other quantities is thereby manifest in the logical form of the nomic equations.

Frank's view nicely accords with our intuition that a body's acceleration is caused by the mass of and net force on the body; although $F = ma$ is not a differential equation, it is equivalent to $F = m\, dv/dt$. Likewise, Fourier's law of heat conduction says that the heat flux density at a given location (the instantaneous rate at which heat energy is flowing across that spot) is proportional to the negation of the temperature gradient there. Frank's account entails that temperature differences cause heat flow, which sounds right.

However, Newton's law of gravity ($F_g = GmM/r^2$, relating the gravitational force F_g between point masses m and M to their separation r), though, does not involve a differential equation. Frank could reply that although some natural philosophers have regarded forces as

real, even directly observable (W. Thomson and Tait 1888: 220; Armstrong 1997: 212), forces are actually unreal; talk of them is merely a convenient device for tallying various causal influences (such as distant masses or local fields). Since 'F' stands for nothing real, it is disqualified from representing either cause or effect, and we should turn our attention from F_g = GmM = r^2 to a_g = GM = r^2. This is a differential equation in disguise (since a = dv/dt), so Frank's account interprets the body's acceleration as the effect of a distant mass.

However, consider the Lorentz-force law: that the magnetic force F_m on a point charge q moving with velocity v in magnetic field B equals (q/c) $v \times B$. Intuitively, the force is caused by q, v, and B. But the only instantaneous rate of change in the equation is v (which is the time rate of change of the body's trajectory). Accordingly, Frank's account must incorrectly deem v to be the effect. If we replaced 'F_m' in the equation with 'ma', then the equation would contain two instantaneous rates of change, and Frank's account cannot designate which is the effect.

Other laws also pose difficulties for Frank's proposal. Faraday's law of electromagnetic induction is that curl$E$ = $(1/c)$dB/dt. Presumably, then, Frank's account entails that an electric field's 'rotation' at a given point causes the magnetic field's rate of change there. However, Faraday's discovery was that a changing magnetic field (or perhaps the changing current that causes the changing magnetic field) causes a(nother) current to flow; it is the electric field that is 'induced'.

3. LOCALITY

Rather than trying to discern causal relations from a law's syntax, most fans of causation prefer to ground causal relations in scientifically respectable facts that go beyond the bare nomic equations, such as facts expressed by counterfactual conditionals or about the flow of conserved quantities. However, besides finding no causal relations in the nomic equations, Russell has another reason for interpreting classical physics as having discovered that there are no causal relations.

Many philosophers, notably Hume (1739/1978: 75), have maintained that an essential feature of causal relations is their 'locality': an event's causal influence cannot jump across a spatial or temporal gap. In other words, its causal influence cannot be felt at a location distant in space or time without being felt throughout a continuous path traversing the intervening region. However, Russell could argue, classical physics outgrew locality. If there are causal relations, then a body's velocity at time t_2 is the cumulative effect of the body's instantaneous accelerations between t_1 and t_2 coupled with $v(t_1)$. But the body's acceleration (its velocity's instantaneous rate of change) at time T cannot affect the body's velocity at T; time must pass before the velocity changes as a result of $a(T)$. (Suppose that for a long while, a body is at rest with zero acceleration. Then from T onwards, it feels a constant non-zero net force and so possesses a constant non-zero acceleration. At T, its speed has not yet changed; it is still zero.) But if the body's $a(T)$ has no effect on the body at T, then when does it first have an effect? For any time t later than T, there is some finite interval of time between T and t. So if the first moment at which the body's velocity is affected by $a(T)$ is some such t, then there is a finite

temporal gap between cause and effect, contrary to locality.

The best response to this argument is to deny that there is a first moment at which the body's velocity is affected by $a(T)$. There is only T: the last moment before $a(T)$ has an effect on the body's velocity. Locality is then satisfied, since $a(T)$ has an effect at some time t later than T only if $a(T)$ has effects at every moment in between.

This conception of locality can be expressed more explicitly (Lange 2002: 13):

> Temporal locality (TL): For any event E and for any finite temporal interval t > 0, no matter how short, there is a complete set of E's causes such that for each event C (a cause) in this set, there is a moment at which it occurs that is separated by an interval no greater than τ from a moment at which E occurs.

If T^{TM} were the first moment after T at which the body's velocity is affected by $a(T)$, then there would presumably be no complete set of $v(T')$'s causes within τ $(T' - T)$ of T'. But in classical physics, for any τ > 0, no matter how short, the body's $v(t_2 - τ)$ and instantaneous accelerations in $[t_2 - τ, t_2>$ form a complete set of $v(t_2)$'s causes, satisfying TL. This set is 'complete', despite omitting some causes of $v(t_2)$ preceding $(t_2 - τ)$, because it supplies a complete causal explanation of E. An event may have many 'complete' sets of causes, just as my two parents form a 'complete' set of my causal ancestors and my four grandparents do too. (Three would be incomplete.)

But TL, even combined with its spatial analogue SL, is insufficient to capture locality, since locality requires that spatial and temporal gaps between cause and effect be capable of being diminished arbitrarily *together*. If there are causes arbitrarily near in time to E, but those nearer in time to E are farther in space from E, and there are other causes arbitrarily near in space to E, but those nearer in space to E are further in time from E, then SL and TL may be satisfied without satisfying:

> Spatiotemporal locality (STL): For any event E, any finite temporal interval $τ > 0$, and any finite distance $δ > 0$, there is a complete set of causes of E such that for each event C in this set, there is a location at which it occurs that is separated by a distance no greater than $δ$ from a location at which E occurs, *and* there is a moment at which C occurs *at the former location* that is separated by an interval no greater than τ from a moment at which E occurs *at the latter location*.

STL entails TL and SL, but not vice versa.

However, Russell could argue that according to classical physics, the universe violates these locality principles. Laws associate some gravitational forces on Earth with matter on the Sun, millions of miles away; these forces appear to violate SL. Newton famously refrained from offering any hypothesis regarding the cause of gravitational forces ('hypotheses non fingo'). Some natural philosophers (such as Huygens and Leibniz) regarded gravitational action at a distance as metaphysically repugnant. But (Russell could say) nature refused to gratify these

metaphysical prejudices. Newton's successors found laws of electromagnetism correlating a charge at one moment with an electromagnetic force on another charge at a later moment some distance from where the first charge was (the time delay in a vacuum being roughly 1 second for every 3×10^8 metres between them). Evidently, electromagnetic forces obey neither SL nor TL. Even 'collisions' between billiard balls (that paradigmatic causal relation) actually involve charges repelling one other electrostatically across a small spatial gap rather than incompressible bodies touching one other. The world is not mechanical and hence not causal.

One possible response to this argument is to deny that locality is essential to causation. Though Newton himself deemed action-at-a-distance an 'Absurdity' (Cohen 1978: 302–3) and Clarke called it a 'contradiction' (Alexander 1956: 53; cf. Hertz 1893: 122–3), it is difficult to find an argument (that refrains from begging the question) showing that a cause cannot act where or when it does not exist.

Alternatively, a force may be interpreted as having *local* (though non-mechanical) causes: the affected body's charge and the corresponding *field* at the body's spatiotemporal location. Maxwell (1873/1954: ii. 493) argued that electric and magnetic fields exist since they possess energy (and momentum). By assuming that individual parcels of energy travel along continuous paths, Poynting (1884) found an expression for the energy flux density (the rate and direction of energy flow at each location) in regions empty of matter but containing electric and magnetic fields. However, Poynting's solution yields results that Hertz (1893), Jeans (1932: 129), and J. J. Thomson (1886: 151–2), among others, found counterintuitive (such as that energy circulates around a stationary magnet near an unmoving charge). Moreover, it is not evident that energy comes in parcels with identities over time; no parcel can be marked in order for its trajectory to be tracked. Furthermore, Maxwell's equations radically underdetermine energy flow; there are infinitely many ways besides Poynting's to depict energy as moving continuously through space devoid of matter. Accordingly, Jeans suggested that energy is merely a theoretical device. It then could not underwrite the reality of fields. This issue is not resolved within classical physics.

4. DETERMINISM OR UNCAUSED EVENTS?

To defend causal eliminativism, Russell might have argued that a body is said to require a cause to accelerate, but not to persist in its speed and direction. This distinction between natural and forced motion seems like a vestige of modelling causation on one agent making another do something that she would otherwise not have done (Sellars 1963: 13–14). Rather than depict some events but not others as caused, Russell recommends that we dispense with causes and instead treat all motion identically: as determined by law and initial conditions.

There is another alternative to depicting forced motions as caused but natural motions as uncaused: to regard natural motions as caused by the absence of net force (which may involve the presence of balanced component forces). However we resolve this case, though, it has recently been argued that even determinism fails in classical physics: Newton's laws of motion and gravity permit events (involving 'forced' motion) to occur uncaused. Newton's

laws enable a closed system of point masses accelerating under their mutual gravitation to undergo an infinite number of triple near-collisions in a finite time (as the sequence of encounter times converges to some particular moment). By the slingshot effect resulting from these close approaches, certain bodies attain infinite acceleration in finite time (their kinetic energy drawn from the system's gravitational potential energy) and so afterwards are absent from any finite region of the universe. They are literally nowhere to be found. Since Newton's laws are time-reversal invariant, they permit this sequence of events to proceed in reverse, so that 'space invaders' suddenly appear in the system from nowhere—uncaused and violating determinism (Earman 1986).

Likewise, Laraudogoitia (1996) considers infinitely many spheres of unit mass at rest, lined up at $x = 0$, ½, ¾, ⅞, ..., each mass half the diameter of its predecessor. At $t = 0$, the leftmost body is struck by an identical body moving to the right with unit speed. Each body collides elastically with the body to its right, coming to rest and replacing it in line. Hence, by $t = 1$, each body has been brought to rest by another. Since Newton's laws are time-reversal invariant, they permit this sequence of events to proceed in reverse: while the bodies could remain at rest forever, they could also spontaneously begin to move, finally firing a body leftward. Each body would have been caused to move by another, but nothing would have initiated the ripple.

As another example, Norton (2007) considers a symmetric frictionless domed surface in a downward gravitational field (with gravitational acceleration g); at distance r from the apex as measured along the dome's surface, the dome's height is $(2/3g) r^{3/2}$ lower than at its peak. Hence, at any point on the dome, a unit point mass there would undergo an acceleration tangential to the surface of $g(\mathrm{d}h/\mathrm{d}r) = r^{1/2}$. Since the acceleration is zero at $r = 0$ (the apex), a unit mass-point placed there at rest at $t = 0$ accords with Newton's second law by remaining there. However, the body also accords with Newton's second law by remaining at the apex until some arbitrary time T, after which its separation from the apex is $(1/144) (t - T)^4$. At every moment $t > T$, the body experiences non-zero force and undergoes non-zero acceleration in accordance with Newton's second law, but at T, its last moment at the apex, nothing causes it to begin moving. Conditions at that moment are no different from conditions at earlier moments.

Of course, the space invaders' unanticipated appearance violates mass, energy, and momentum conservation, and the start of Laraudogoitia's ripple violates energy and momentum conservation (and the total mass is infinite). If classical physics is interpreted so austerely as to exclude these conservation laws (Earman 1986: 37–9), then these trajectories are permitted by classical physics. On the other hand, although Newton does not posit these conservation laws explicitly, his physics seems to presuppose some such principles. At least, it is difficult to see how classical physics could make any predictions at all if it left open the possibility of new bodies appearing spontaneously anywhere anytime.

Norton's mass-point case does not violate conservation laws. However, it and the other examples of uncaused events may perhaps be ruled out without invoking conservation laws. Maclaurin (1748/1971: 113) takes something beyond the bare equations to be implicit in Newton's laws:

It is a part of the same law, that a body never changes the direction of its motion, of itself,

but by some external influence only; and it is as natural a consequence of the passive nature of body, as that it never changes its velocity of itself. A body has no self-motive power, or spontaneity, if it was to change its direction, how could it determine itself to any one direction rather than to another?

In this spirit, we might enrich the bare equations with some such principle as this:

Every body at every moment at which it exists has a definite location. No body can leave its location without having non-zero velocity (or acceleration or ...) at the final moment at which it occupies that location (if there is a final such moment). For any closed system of bodies, no body can leave its location without some body in the system having non-zero velocity (or acceleration or ...) at the final moment at which every body occupies its 'original' location.

This principle is violated by Laraudogoitia's case, for instance, where $t = 0$ is the final moment at which every body in the system occupies its original location, yet none has non-zero velocity (or acceleration or ...) at that moment.

This principle also precludes a body's following the trajectory

$$r(t) = \begin{cases} 0 & \text{for} \quad t \le 0 \\ e^{-1/t^2} & \text{for} \quad t > 0. \end{cases}$$

This function is infinitely differentiable and all its derivatives vanish at $t = 0$. So such a body apparently has no possible cause at $t = 0$ to leave its location then.

However, the only possible causes of a body's leaving its location that classical physics countenances are the body's velocity and acceleration (caused by impressed forces), not higher-order derivatives of trajectory. Hence, although the above principle permits the trajectory $r(t) = 6t^3$, since its third time-derivative at $t = 0$ is non-zero, classical physics does not seem to find any cause for such a body to leave $x = 0$ at $t = 0$ (since its velocity and acceleration are zero at $t = 0$). On these grounds, perhaps the above, purely kinematic principle should be strengthened to reflect Newtonian dynamics by having the 'or ... 's removed from it. On the other hand, an austere view of classical physics as nothing but Newton's equations of motion would regard the effort to craft any such principle as trying to force classical physics to conform to deterministic prejudices.

5. ANALYTICAL MECHANICS AND TELEOLOGY

Given the system's initial configuration (the initial positions and velocities of its particles) and final configuration, there are various paths (though configuration space) by which the

system may get from one to the other. These paths may differ, for instance, in the time it takes the system to arrive at its final configuration and in the configurations through which the system passes along the way. Roughly speaking, the Euler–Lagrange 'principle of least action' states that the time integral of the system's total kinetic energy is 'stationary' along the actual path as compared to all sufficiently close possible paths. That is, roughly speaking, the sum of the kinetic energies at all the points along the path actually taken is a minimum, maximum, or saddle point as compared to the sums for similar paths that are not taken. (So 'the principle of least action' does not demand that the action be least among all possible paths or even among all similar possible paths.) Similarly, Hamilton's principle states roughly that of all the possible paths by which the system may proceed from one specified configuration to another in a specified time, the actual path as compared to other possible, slightly different paths makes stationary the time integral of the system's 'Lagrangian' (i.e. the difference between the system's total kinetic and potential energies). A 'possible' path may violate energy conservation and other laws; Hamilton's principle picks out the path demanded by the laws. So to apply Hamilton's principle, scientists must contemplate 'counterlegals': what would have been the case, had the system violated natural laws in certain ways. But a 'possible' path must respect the constraints on the system, which may include a body having to remain rigid or in contact with a certain surface.

The variational principles of analytical mechanics make no mention of forces, instead invoking the system's energy. The explanations they supply specify no efficient causes. Variational principles involve integral equations; they determine the system's trajectory as a whole, rather than point by point.

Explanations using variational principles sound teleological; the system appears to aim at making a certain integral stationary. But then the system's final configuration apparently helps to explain the path that the system takes to that destination; later events help to explain earlier ones. That is puzzling. How does a light ray 'know', at the start of its journey, which path would take less time? How can the light adjust the earlier part of its route to minimize its later path through optically dense regions (where it cannot travel as fast) unless it knows about those distant regions before it sets off?

Some natural philosophers, such as Planck, have suggested that variational principles are more basic laws than Newtonian differential equations, especially considering that unlike Newton's laws, variational equations of the same form apply to any set of variables sufficient to specify the system's configuration. Other natural philosophers, notably Leibniz (1969: 478–9), have embraced both mechanical and teleological explanations as equally fundamental. The most common view, however, has been to reject teleological explanations as a relic of anthropomorphic characterizations of nature, and to regard variational principles as logical consequences of more fundamental, mechanical laws. The variational principles follow from the Newtonian differential equations roughly because the entire path can minimize the integral only if each infinitesimal part does (since otherwise, by replacing that part with another, we would create a new path with a smaller integral), and the minimum for each infinitesimal part reflects the gradiant of the potential there, which is the force. The variational principle thus arises as a byproduct of the relation between the force and an infinitesimal section of the path.

FURTHER READING

Important nineteenth-century physics texts aiming to set physics on a proper conceptual foundation include W. Thomson and Tait (1888), Mach (1893/1960), and Hertz (1894/1956); of course, one must start with Newton (1687/1971). General philosophy of physics texts devoting considerable attention to classical physics include Sklar (1992) and Cushing (1998); Lindsey and Margenau (1936) is a worthwhile older text. Lange (2002) includes extensive discussions of the ontological status of fields and energy, as well as different senses of locality. Hesse (1965) is a fine historical survey of arguments regarding fields and action at a distance; Berkson (1974) is a more popular presentation. Frisch (2005) contains further discussion of various senses of locality along with a fascinating account of the role of auxiliary causal assumptions in classical electromagnetic theory. For more on laws of nature, see Armstrong (1983; 1997), Lange (2000), and Carroll (1994; 2004), which elaborate the key problems and popular options. Diacu and Holmes (1996) is a good historical account of the discovery of apparent violations of determinism in classical physics; Earman (1986) is the central philosophical discussion. Feynman, Leighton, and Sands (1964: ii. 19-9) argue that quantum mechanics depicts light as 'aware' of later parts of its path so as to compensate by adjusting its trajectory during earlier parts of its route. Yourgrau and Mandelstam (1968) is a comprehensive source on variational principles, though Nahin (2004) is a more entertaining introduction to them. Stoltzner (2003) emphasizes various philosophers's efforts to grapple with them. Lange (2005) considers what instantaneous velocity would have to be in order for it to be able to serve as a cause of a body's subsequent trajectory.

REFERENCES

ALEXANDER, H. G. (1956). *The Leibniz-Clarke Correspondence*. Manchester: Manchester University Press,.

ARMSTRONG, DAVID A. (1983). *What Is A Law of Nature?* Cambridge: Cambridge University Press.

—— (1997). *A World of States of Affairs*. Cambridge: Cambridge University Press.

BERKSON, WILLIAM (1974). *Fields of Force*. New York: Wiley.

CARROLL, JOHN (1994). *Laws of Nature*. Cambridge: Cambridge University Press.

—— (ed.) (2004). *Readings on Laws of Nature*. Pittsburgh: University of Pittsburgh Press.

CARTWRIGHT, NANCY (1983). 'Causal Laws and Effective Strategies', in *How the Laws of Physics Lie*. Oxford: Clarendon, 21–43.

COHEN, I. B. (ed.) (1978). *Isaac Newton's Papers and Letters on Natural Philosophy*. Cambridge, Mass.: Harvard University Press.

CUSHING, JAMES (1998). *Philosophical Concepts in Physics*. New York: Cambridge University Press.

DIACU, FLORIN, and HOLMES, PHILIP (1996). *Celestial Encounters: The Origins of Chaos and Stability*. Princeton: Princeton University Press.

EARMAN, JOHN (1986). *A Primer on Determinism*. Dordrecht: Reidel.

FEYNMAN, RICHARD, LEIGHTON, ROBERT, and SANDS, MATTHEW (1964). *The Feynman Lectures on Physics*. Reading, Mass.: Addison-Wesley.

FRANK, PHILIPP (1932/1998). *The Law of Causality and Its Limits*, ed. Robert S. Cohen, trans. Marie Neurath and Robert S. Cohen. Dordrecht: Kluwer.

FRISCH, MATHIAS. (2005). *Inconsistency, Asymmetry, and Nonlocality*. Oxford: Oxford University Press.

HEMPEL, CARL G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

HERTZ, HEINRICH (1893). *Electric Waves*, trans. D. E. Jones. London: Macmillan.

—— (1894/1956). *The Principles of Mechanics*, trans. D. E. Jones and J. T. Walley. New York: Dover.

HESSE, MARY (1965). *Forces and Fields*. Totowa, NJ: Littlefield, Adams.

HUME, DAVID (1739/1978). *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, rev. P. H. Nidditch. Oxford: Clarendon.

JEANS, JAMES (1932). *The Mysterious Universe*. New York, NY: Macmillan.

LANGE, MARC (2000). *Natural Laws in Scientific Practice*. New York: Oxford University Press.

—— (2002). *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass*. Malden, Mass.: Blackwell.

—— (2005). 'How Can Instantaneous Velocity Fulfill Its Causal Role?' *Philosophical Review* 114: 433–68.

LARAUDOGOITIA, JON PEREZ (1996). 'A Beautiful Supertask', *Mind* 105: 81–3.

LEIBNIZ, GOTTFRIED WILHELM (1969). *G. W. Leibniz: Philosophical Papers and Letters*, 2nd edn., trans. and ed. L. E. Loemker. Dordrecht: Reidel.

LINDSAY, ROBERT, and MARGENAU, HENRY (1936). *Foundations of Physics*. New York: Wiley.

MACH, ERNST (1893/1960). *The Science of Mechanics*, trans. Thomas McCormack. LaSalle, Ill.: Open Court.

MACLAURIN, COLIN (1748/1971). *An Account of Sir Isaac Newton's Philosophical Discoveries*. Hildesheim: Olms.

MAXWELL, JAMES CLERK (1873/1954). *A Treatise on Electricity and Magnetism*. New York: Dover.

NAHIN, PAUL (2004). *When Least Is Best*. Princeton, NJ: Princeton University Press.

NEWTON, ISAAC (1687/1971). *Sir Isaac Newton's Mathematical Principles of Natural Philosophy and His System of the World*, trans. Andrew Motte, rev. Florian Cajori. Berkeley: University of California Press.

NORTON, JOHN (2007). 'Causation as Folk Science', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Oxford University Press, 11–44.

POYNTING, JOHN HENRY (1884). 'On the Transfer of Energy in the Electromagnetic Field', *Philosophical Transactions of the Royal Society of London* 175: 343–62.

RUSSELL, BERTRAND (1912–13/1953). 'On the Notion of Cause', *Proceedings of the Aristotelian Society*, 13: 1–26; repr. in Herbert Feigl and May Brodbeck (eds.), *Readings in the Philosophy of Science* (1953). New York: Appleton-Century-Crofts, 387–407.

SALMON, WESLEY (1989). 'Four Decades of Scientific Explanation', in Philip Kitcher and

Wesley Salmon (eds.), *Scientific Explanation*, Minnesota Studies in the Philosophy of Science 13. Minneapolis: University of Minnesota Press, 3–219.

SELLARS, WILFRID (1963). *Science, Perception, and Reality*. London: Routledge.

SKLAR, LAWRENCE (1992). *Philosophy of Physics*. Boulder: Westview.

STOLTZNER, MICHAEL (2003). 'The Principle of Least Action as the Logical Empiricist's Shibboleth', *Studies in History and Philosophy of Science* B 34: 285–318.

THOMSON, J. J. (1886). 'Report on Electrical Theories', *BAAS Report—1885*. London: John Murray, 97–155.

THOMSON, WILLIAM, and TAIT, PETER GUTHRIE (1888). *Treatise on Natural Philosophy*. Cambridge: Cambridge University Press.

YOURGRAU, WOLFGANG, and MANDELSTAM, STANLEY (1968). *Variational Principles in Dynamics and Quantum Theory*. Philadelphia: W. B. Saunders.

CHAPTER 32
CAUSATION IN STATISTICAL MECHANICS

LAWRENCE SKLAR

1. CAUSATION AND STATISTICAL MECHANICS

In a well-known historical development many positivists offered accounts of explanation that eschewed the introduction of causal notions. Partly motivated by Hume's eliminative account of the causal relation, and partly motivated by the apparent appearance within legitimate science of explanations that didn't seem causal in nature, they offered accounts of explanation as, essentially, subsumption of particular occurrences or features under fully lawlike or statistical regularities. Then came all the arguments designed to convince us that this could not be the full story. In particular causal notions must be introduced, it is claimed, to distinguish between regularities that are genuinely explanatory and those that are not.

The main importance of statistical mechanics for philosophical explorations of causality is, of course, the allegation that at least one aspect of the causal relation, its temporal asymmetry, can be grounded in the deep temporal asymmetry of processes in nature that is summed up in the time asymmetric Second Law of Thermodynamics and in the attempted explanations within statistical mechanics of temporally asymmetric entropic increase. Since this topic is explored in Ch. 20 above, I will not pursue this deep issue here.

Here, instead, I will briefly explore two ways in which explanations in statistical mechanics seem to have aspects that are not of a straightforward causal nature. The familiar pattern of explanation we find in physics is well known: find means of characterizing momentary states of systems. Find lawlike generalizations that allow one to infer future states of a system from earlier states. Take this pattern to be a representation of the way in which earlier momentary states of a system causally determine its future conditions.

Within statistical mechanics this lawlike connection of state to state that is, intuitively, a representation of a causal relation, most certainly does play an ineliminable role. In statistical mechanics it appears at the micro-level as the postulation that the full state of a system at one time can be specified by the dynamical state of all its micro-constituents (the positions and momenta of the molecules in a gas or, alternatively the wave function of these at one time), and that this state at one time generates, following the laws of dynamics (classical or quantum) the future dynamical state of the system characterized in these micro-constituent terms.

So what is 'non-causal' in nature in explanations in statistical mechanics? I will explore two issues: (1) The peculiar 'transcendental' nature of explanation in equilibrium theory in statistical mechanics; (2) The need for introducing some a priori probability posit over initial

conditions of systems in non-equilibrium theory.

2. EQUILIBRIUM THEORY

Long before the introduction of the principle of energy conservation into thermodynamics in the form of the First Law, and long before the introduction into thermodynamics of a general principle of time asymmetry of processes by means of the definition of entropy and the postulation of the Second Law, thermodynamics had its beginnings in the phenomenological principles governing simple equilibrium systems. First came the realization that pressure and volume were inversely related for gases in equilibrium, and later, with the introduction of temperature into the picture, came the famous ideal gas law, $PV = kT$. These regularities, governing the relations between a few macroscopically measurable physical parameters of simple systems, were the first indication that a realm of lawlike behaviour existed that eventually became thermodynamics. Interestingly, these relations are paradigms of the kind of lawlike regularities that don't appear at first inspection to be causal in nature. For one thing, they specify relations among quantities at a single time, and they say nothing about how quantities developed or were brought about from the previous states of the world.

But why do these simple relations hold? Here we must look at how the phenomenological equilibrium thermodynamics of simple systems became embedded into equilibrium statistical mechanics at the end of the nineteenth century and the beginning of the twentieth.

The standard treatment of equilibrium in statistical mechanics looks like this: one considers a system characterized by some macroscopic constraints. One takes the system to be composed of numerous microscopic components whose interaction is given by some standard dynamics. Next one imposes over the possible dynamical states of the system, characterized in microscopic terms, a standard probability distribution whose nature depends on the type of constraints to which the system is subject (energetic isolation vs. thermal contact with a constant temperature heat bath, for example). Using the probability distribution imposed, one calculates average values for some functions of the dynamical micro-states of the system. These mean values are 'identified' with the macroscopic parameters of the system. Out of this one hopes to derive the well-known equilibrium relations among the macroscopic parameters of the system.

The intuitive rationale behind it all is something like this: a system will move from one microscopic dynamical state to another. Over a long time the system will, for overwhelmingly dominant time periods, have its microscopic state in a class of states that correspond to the system being near the equilibrium condition. So the 'overwhelmingly most probable' micro-condition is the equilibrium condition. But for systems of vast numbers of particles, and for appropriate functions of the micro-conditions, the overwhelmingly most probable value of such a function will agree with its mean value over all possible micro-conditions for the system. This is an argument from consideration of what is called the 'thermodynamic limit' and has been explored at great depth in the literature.

Justifying the claim that this picture represents how systems really behave is the province of non-equilibrium theory. Only in that broad context could we find some rationale for the

claim that systems will indeed spend most of their life in micro-states corresponding to macroscopic equilibrium. But there is a curious autonomous equilibrium theory whose explanations I want to focus on here.

Why should we take the standard probability distribution as the right one for calculating the mean values of micro-condition functions that we are going to identify with macroscopic parameter values? After all, there are innumerable probability distributions that can be assigned over a collection of possible microscopic states for a system. A priori rationales for choosing the standard distribution have been proposed, but they are extremely dubious. The standard probability measure is one uniformly distributed over the micro-states. Can't we then use some 'ignorance' argument to justify it application? Here the problem is the familiar one that 'uniformity' of a probability distribution is relative the characteristics with which one chooses to describe the micro-states. Uniformity with respect to a position-momentum characterization, which is the probability distribution actually chosen and which 'works', is not the same as uniformity with respect, say, to a position-energy characterization of the states.

Ergodic theory attempts to answer this question in a curiously autonomous way that ignores the question of why systems go to equilibrium states and questions about how they get there. The argument, rather, goes like this: Equilibrium is a condition of a system that is supposed to be unchanging in time. So if we are to associate the macroscopic parameters of equilibrium with something calculated using a probability distribution over the microscopic states possible for the system, then that probability distribution must be one which is time invariant. Suppose we put a probability distribution over the microstates that assigns a certain probability to some collection of microstates. A system in any microstate will, following the causal structure of the dynamics, later be in some different microstate. So a given collection of microstates specified by some constraint will have systems leaving that collection and systems entering that collection as time goes on. For a probability distribution to be invariant in time, the 'volume' of systems leaving a collection in any time interval must always be balanced by the volume of systems entering that collection in that time interval.

It is easy to show that the standard equilibrium probability distributions have this time invariant character. But are they the *only* such time invariant probability distributions? This question was posed early in the development of the theory, and only (partially) answered after many decades and many false starts. What one can show is that for some idealized models that seem relevant to realistic characterizations of actual systems, there is only one probability distribution that is both invariant in time and which is such that the probability distribution assigns zero probability to any set of conditions to which the standard distribution assigns zero probability. And that is the standard distribution.

Whether or not this is a sufficient autonomous rationale for the usual equilibrium theory of statistical mechanics is a subject of much debate. But that is not our concern here. Here the point is that the mode of explanation being invoked is curiously non-causal. And it is a kind of explanation that appears in a variety of contexts within physics.

In autonomous equilibrium theory no attempt is made to explain why systems approach the equilibrium state. Nor is there any attempt to describe and explain the path by which systems approach equilibrium. Instead the existence of equilibrium states is just taken as a posit. Furthermore the assumptions are made that rationalize the use of a probability distribution

over the possible micro-conditions of the system and that posit taking mean values calculated by means of this probability distribution as appropriate representatives of the observed values of macroscopic parameters characterizing the equilibrium state. Here again the picture of the system as existing for dominant periods of time at or near equilibrium microscopic conditions is one that could only be justified within the non-equilibrium theory. The real causal story behind equilibrium is then, put to the side in the autonomous equilibrium approach.

Instead the autonomous equilibrium theory concerns itself with a 'transcendental deduction' of the appropriate probability distribution to be invoked, asking which such probability distributions could have the time invariance appropriate for being used to characterize an assumed time invariant equilibrium condition. Within this explanation causation plays its role, since the demonstration of temporal invariance of the standard probability distribution, and the partial demonstration of the uniqueness of the standard probability distribution as the only temporally invariant such distribution, rely upon the causal dynamics of the micro-conditions as expressed by the standard laws of dynamic evolution. But the overall explanatory pattern is not the familiar explanation of one condition by the eliciting of its causal development.

This kind of 'explanation by transcendental deduction' is not confined to just one case in physics. Here is another example: we know from our observational experience that materials cooled sufficiently will often turn solid as crystals. Here an astonishing transition has occurred. Before crystallization, each atom or molecule of the substance was 'aware' only of its near neighbours' conditions, aware by means of the forces the near neighbours exerted on it due to their relative positions and orientations. But once crystallization has occurred, every atom or molecule has its position rigidly fixed relative even to the most distant atoms or molecules in the sample. Order has transformed from short range to long range. An exactly similar situation describes the onset of ferromagnetism, where the spins of atoms previous to the onset of ferromagnetism are correlated only with those of near neighbours by spin–spin interactions. But after the onset of ferromagnetism every atom has its spin aligned with even the most distant other atoms in the sample.

The full explanation of why systems crystallize or become ferromagnetic, and by what path they do so, requires a full-scale non-equilibrium theory. But one can make astonishing explanatory progress by ignoring such issues, positing that such states of long-range order do exist, and drawing inferences from the very fact that they do. Just as the fact that time invariance characterizes equilibrium allows us to make inferences about just what the appropriate probability distribution must be like over the micro-states if it is to be used to describe equilibrium, the facts about the long-range order of crystals and ferromagnets allow us to make inferences that allow characterizations of features of these states and the transitions to them.

It is from theories of this kind, generally called renormalization group explanations, that one can understand why it is that there are deep formal similarities between the features that characterize transitions of systems to states of long-range order, such as the shape of curves of specific heats that describe the transition, that are 'universal' in nature, characterizing a wide class of systems whose members can differ substantially in the specifics of the dynamics that holds between their micro-constituents. It is this theory that lets us understand why only such features as the dimensionality of the system and the degrees of freedom available to its constituents, along with very general features of the interaction forces among the components,

are all that matter in determining the structures of the transition to long-range order.

Once again, causal considerations certainly function in these explanations, but the overall form of the explanation is not one that elicits the causes of long-range order, but one that derives features of this order and of the transition to it from the fact, known from empirical experience, that such long-range orders and the transitions to them do, in fact, exist.

3. NON-EQUILIBRIUM THEORY

Systems prepared in non-equilibrium conditions go to equilibrium. Then they stay in the equilibrium state. This is the most fundamental fact of the thermodynamic behaviour of systems in the world. Why does this happen?

Of course, it really doesn't really happen quite that way. From the statistical mechanical point of view inevitable, monotonic approach to equilibrium and permanence of the equilibrium condition are not the case. All the claims and posits must get the appropriate modifications to deal with the statistical and probabilistic nature of things at the underlying level. So we may adopt the Boltzmann–Ehrenfest picture of a collection of systems, each one of which spends most of its eternal life at or near equilibrium, but with fluctuations away from the equilibrium condition with the larger fluctuations being the rarer. Or we might, instead, think in terms of the Boltzmann–Lanford picture of a collection of systems prepared in non-equilibrium with an overwhelmingly high probability of the systems following a monotonic approach to equilibrium.

We would like to describe the path taken in the approach to equilibrium of a non-equilibrium system. In a very few cases we can do this. For rare gases, for example, there is the Boltzmann equation or the equivalent Maxwell transfer equations. In a few other cases other 'kinetic equations' can be derived. For the most part, though, physics has been stymied in trying to find kinetic equations. Even for moderately dense gases science is at an impasse.

It is hardly surprising that causation as intuitively understood plays a significant role in the explanation of non-equilibrium behaviour. All derivations of the kinetic equations rely on tracing the interactions of the micro-components of the system as they interact with one another. Typical, for example, is the picture of the ideal gas as composed of molecules that move freely through space except when they collide elastically with one another or with the walls of a confining box.

But bringing into account the dynamical laws governing the motion and interaction of the micro-constituents won't do by itself. Something needs to be done to take account of the fact that, from the causal perspective, how the system will evolve depends upon its *initial conditions*. Early statistical mechanics was plagued by the fact that the derivations of the kinetic equations relied upon a process of continually 're-randomizing' the initial state of a system at every moment of time, even though the proof of the legitimacy (or even consistency) of this procedure was lacking.

Since the early days of the theory there have been several attempts at ridding it of the necessity of a re-randomization posit. One approach follows the Boltzmann–Ehrenfest interpretation and seeks to show how the dynamics causes a continual stirring up of any

ensemble of systems so that 'in the temporal limit' any ensemble of systems will come to look 'coarsely' like the equilibrium ensemble for the system's constraints. Another approach tries to show a 'high probability of approach to equilibrium' of any system started in an initial ensemble.

What is essential here is that in both these approaches one cannot get anywhere with trying to account for the behaviour of systems without introducing some probabilistic posit over the possible initial conditions available to the system when it is initially prepared in its non-equilibrium condition. The 'right' posit to impose is easy to find. It is just the usual uniform distribution of probability over the set of possible initial conditions where uniformity is relative to the usual position-momentum characterization of the space of initial conditions.

The most notorious problem within non-equilibrium statistical mechanics is that of time asymmetry. The underlying dynamics of the micro-constituents is time symmetric. It is important to note here that some way of introducing time asymmetry into the system, by arguing that the dependence of states of the micro-constituents on one another being causal has a 'built-in' time asymmetry to it, is not often taken seriously as the right way to get time asymmetry into the theory within physics. The almost universal assumption is that any time asymmetry of causation is to be accounted for by the time asymmetry of the world explained by the theory, and not the other way around, although in early statistical mechanics time asymmetric 'frictions' were frequently speculated upon. We shall soon note that a kind of lawlike time asymmetry does play a role in some recent speculations.

The usual probabilistic posit over initial conditions also makes no reference to time order. But it is invoked in such a way as to introduce time asymmetry into the theory, being applied only to the *initial* and never to the *final* states of the micro-constituents of the system. But the deep question remains as to *why* the posit is appropriately differentially applied in this way.

The most common 'way out' for introducing time asymmetry into the theory is to posit a grand cosmic time asymmetry as explanatory of the thermodynamic time asymmetry of individual systems. Although some accounts rely upon the time asymmetric fact that our known cosmos is expanding into the future, the most widely adopted view is that the time asymmetry of the cosmos is to be grounded, rather, on a posited, particularly low entropy, highly regular, initial state of the cosmos just after the Big Bang initial spatial singularity from which the expansion took off. Whereas most models posit a high entropy initial state for the matter of the universe, in the sense of assuming a kind of equilibrium temperature uniformity for it, the low entropy is imposed by assuming that at the early time matter had a uniform distribution in space. Due to the purely attractive nature of the gravitational interaction, this is an extremely low entropy state. The idea is that as matter coalesced into a lumpy configuration of hot stars and cold empty space, one had a *lowering* of the thermal entropy paid for by an increase in the entropy of the spatial distribution of matter.

Many questions remain unanswered. Some have to do with the question of why the initial condition had such low entropy, given the (alleged) high probability of high entropy states. One answer invokes a new lawlike constraint (All white holes have low entropy—Penrose). Another answer posits multiple universe solutions with 'anthropic' arguments added (there are lots of universes, ours is the rare kind, but only in the rare kind will observers exist). A third response is to the deny that there is any sense in talking about the probabilities of whole universes (cf. Hume on the teleological argument and universes not being as plentiful as

blackberries).

Other deep questions concern the issue of how to go from a posit of initial low entropy for the universe as a whole to the entropic behaviour of individual, small systems temporarily more or less thermally isolated from the rest of the world. After all, it is the fact that these so-called 'branch systems' have their entropies increase parallel to one another in a single time direction that we call the future that is the subject of the classic Second Law of Thermodynamics and that receives the probabilistic surrogate of that law in the statistical physics of non-equilibrium.

The positing of initial low entropy for the universe as a whole counters the paradox that statistical mechanics would seem to entail that, with overwhelmingly high probability, the cosmos as a whole is in a near equilibrium state. But the actual cosmic world is in a highly non-equilibrium state both at the beginning of a time period for which a branch system is isolated and at the end of that period as well. By itself, then, cosmic non-equilibrium won't suffice to explain the parallelism in time of the entropic increase of branch systems.

One could simply add to one's basic theory a posit to the effect that, in the probabilistic sense, the entropic change of branch systems will parallel in time the entropy increase of the cosmic system explained by its initial low entropy and the usual time-neutral probabilistic posits over initial conditions. Not only would such an account be 'non-causal' it would seem to close investigation into all the real explanatory questions. One could also just posit something to the effect that for branch systems it is legitimate to use the usual probabilistic posit over initial conditions at the time origin of the system's separation from the cosmic whole, but not at the final moment prior to the system rejoining the whole. Once again, though, we would have an account that is not only non-causal in nature, but seems to simply restate the problem rather than solving it.

A more promising approach uses an analogy with a well-known result in equilibrium statistical mechanics. Consider an energetically isolated system at equilibrium. Think of it as being made of numerous subsystems considered as energetically isolated from one another for a short interval of time. Then if we use the usual probabilistic posit, these systems will be distributed in a 'canonical' distribution, with their temperatures fluctuating about the temperature of the equilibrium system as a whole.

Now consider the non-equilibrium case of a cosmos as a whole starting from an extremely low entropy initial state and going towards equilibrium. Consider it as being made up of a vast number of energetically isolated subsystems. Impose the usual, time-neutral, probabilistic posit over initial conditions. It is likely that one can derive a result that the overwhelmingly most likely situation is of the increase of entropy of the cosmic system as a whole being accounted for by entropy increase taking place in the vastly overwhelming large majority of branch systems. This is contrasted with, say, a few branch systems picking up the bulk of the entropy increase while most of the systems suffer little or no entropy increase, or even have entropy decreases, in the 'future' time direction given by the entropy increase of the cosmic system.

Such an account would be problematic. It still would have the dramatically non-causal element of the need to make a probabilistic posit over initial conditions which has no causal explanation. It would also be problematic since what we want to explain is the parallel increase in time of ordinary branch systems, not of pieces of the cosmos that remain

energetically isolated from one another for all time.

Proposals to resolve the problem by invoking some primitively time-asymmetric notion of causation do not seem plausible as solutions either. Even if we simply take time order as a given, rather than having time asymmetry somehow reduce to the entropic change asymmetry of systems, and even if we insist that physical states cause only future states and not past states, it is hard to see how any of that will provide a basis on which to establish the appropriateness of positing the probability of initial states of systems according to the standard rules. One problem with such proposals is the existence of systems whose entropic behaviour can be constrained to go in the 'wrong' temporal direction. That would seem to be impossible if the claim were correct that the entropic increase normally experienced were the direct result of some underlying, inviolable causal direction from past to future.

Other proposals are continually floated that try to explain either on the basis of some primitive causal asymmetry or on the basis of some primitive epistemological asymmetry the reason for imposing on initial states a time-asymmetric probability distribution that holds over initial but not over final states. Such proposals can be found, for example, in Gibbs, Watanabe, and Schrödinger. But they have usually been met with a scepticism familiar since the important Ehrenfest critique of statistical mechanics of 1910.

One recent proposal would make the needed explanatory scheme look causal, modulo the posited initial cosmic low entropy which still seems, in most accounts, not to have any causal explanation. To solve the great measurement problem mystery of quantum mechanics, Ghirardi, Rimini, and Weber, and others have proposed that in addition to the usual quantum state function there is a stochastic element that acts on all fundamental particles of the world. This additional element gives a certain, tychistic, physical probability that in any given time the state function of a single particle will collapse to a near eigenstate of position. The stochastic element operates multiplicatively on many-particle wave functions, so that collapses to position eigenstates are improbable for single particles or systems of few interacting particles, but are highly probable in short time intervals for large systems.

There is little to support this GRW account—other than the fact that it is one of the very few intelligible accounts that does explain how superposition states can become eigenstates upon measurement! However Albert has recently suggested that were the GRW account the true solution of the measurement problem in quantum mechanics, it could play a role in the solution of the branch system problem in statistical mechanics as well. Here the basic idea is that a kind of randomization that propagates into the future, but not into the past, is provided to systems by the constant impinging upon their quantum states of the underlying, forward propagating, stochastic influence of the probability fields postulated by the GRW account of quantum measurement.

Such an account will have many subtleties. For some systems the GRW interference is postulated to take place during the evolution of the temporarily isolated system. For other systems an alternative account is needed. Spin-echo systems, for example, show a kind of thermodynamic entropy increase when left alone, but retain their internal quantum correlations among the components of the system as evidenced by the possibility of a 'Loschmidt demon' reversal of the system into showing apparently anti-thermodynamic behaviour. For such systems the Albert account requires that the GRW interference be posited not as acting on the evolution of the system, but by accounting for the 'random' nature of the

environmental initial conditions of the originally non-isolated system. Other puzzles arise when one considers systems with few components (hence low probabilities of GRW interference) but which we still expect to show thermodynamic approach to equilibrium.

So this account remains an interesting speculation. There is still no experimental data to convince us that GRW is the correct account of the quantum measurement collapse of the wave packet. Such data is conceivable, since the theory does predict, contrary to quantum mechanics, wave packet collapses of perfectly isolated systems in superposition states. And there is still no detailed account dealing with actual magnitudes that would convince us that were GRW true at the quantum level, it could be this stochastic element of the world that accounts for the time-asymmetric, parallel in time, approach to equilibrium of branch systems whose behaviour parallels that of a cosmos started in a grossly non-equilibrium condition.

But were this account the correct one, it would provide a 'causal' account of at least one primary part of the general explanation of non-equilibrium behaviour. Of course the kind of causation would be 'stochastic causation' and not what was once taken to be the core of causation, that is, determinism.

4. SUMMARY

So to summarize:

1. In equilibrium statistical mechanics we find explanations that are not causal in their nature. They work by assuming a fundamental fact about the universe (in this case the existence of equilibrium states and the posited micro-structural and statistical account for them), and then ask how such posited states 'are possible'. Such 'how possible' explanations can be called, following Kant, 'transcendental deductions' within physical science. In the case of equilibrium statistical mechanics what is 'deduced', (or, rather, quasi-derived) from the posited state—equilibrium—is the usual standard probability distribution over initial conditions of traditional equilibrium statistical mechanics.

2. In non-equilibrium statistical mechanics we expect to find explanations more plausibly called 'causal'. Yet even here there are puzzles. First of all, the standard account requires positing a special initial condition for the cosmos at the Big Bang. No matter what state we posited at 'time zero' it would seem peculiar to ask for a 'causal' explanation of that state. Unless, of course, we enter the realm of wild speculation positing 'temporally earlier' states in super-cosmic realms—whatever that might even mean. Of course what bothers those working in statistical mechanics the most about the initial Big Bang posited state is that we must posit a state which by the basic probabilistic principles of statistical mechanics is 'wildly improbable'.

Second, even given the posited low-entropy initial state of the cosmos, a full account of non-equilibrium behaviour requires a demonstration that temporarily isolated 'branch systems' will have (probabilistically) their entropies increase parallel in time to one another and to the entropy increase of the cosmic system. The derivation of this standard requires positing the usual probability distribution over *initial* states of isolated systems, but not over their final states. Since causal accounts usually leave choices of initial conditions open, it is

hard to see how such an asymmetric posit of probabilities over initial conditions can be 'causally' accounted for within the standard theory.

Although a recent proposal would offer a causal, but non-deterministic, account of this time asymmetric distribution over the micro-states of systems, such an account remains speculative both with regard to its needed posit of a stochastic level of the universe beyond that treated of in quantum mechanics, and in the suitability of such an account to do justice to the needed probabilistic posits of non-equilibrium statistical mechanics.

FURTHER READINGS

A survey of the basic structures of equilibrium theory in statistical mechanics, including the role of ergodic theory and of the thermodynamic limit can be found in Sklar (1993: ch. 5). For an introduction to the physics and mathematics of renormalization group approaches, see A. Bruce and D. Wallace, 'Critical Point Phenomena', in Davies (1989: ch. 8). For a thorough philosophical exploration of a variety of explanatory structures in physical theories that have 'transcendental' aspects see Batterman (2002). An exposition of the variety of methods used in non-equilibrium statistical mechanics as well as a philosophical discussion of these can be found in Sklar (1993: chs. 6, 7). A seminal work on the role of cosmology in grounding the time asymmetry of statistical mechanics as well as a deep discussion of the branch system problem is in Reichenbach (1956: ch. 3). For a more recent discussion of some of the problems of grounding the parallelism in time of entropic increase of branch systems see Sklar (1993: ch. 8). A concise formulation of the Gibbs–Schrödinger–Watanabe approach is in Gibbs (1960: n. 190). For an introduction to the GRW approach to solving the measurement problem in quantum mechanics, and for an exposition of the possibility that this GRW field could perhaps solve the branch system problem in statistical mechanics, see Albert (2000: ch. 7).

REFERENCES

ALBERT, D. (2000). *Time and Chance*. Cambridge, Mass.: Harvard University Press.
BATTERMAN, R. (2002). *The Devil in the Details*. Oxford: Oxford University Press.
DAVIES, P. (1989). *The New Physics*. Cambridge: Cambridge University Press.
GIBBS, J. (1960). *Elementary Principles in Statistical Mechanics*. New York: Dover.
REICHENBACH, H. (1956). *The Direction of Time*. Berkeley: University of California Press.
SKLAR, L. (1993). *Physics and Chance*. Cambridge: Cambridge University Press.

CHAPTER 33
CAUSATION IN QUANTUM MECHANICS

RICHARD HEALEY

1. INTRODUCTION

There is widespread agreement that quantum mechanics has something radical to teach us about causation. But opinions differ on what this is.

Physicists have often taken the central lesson to be that many physical events occur spontaneously, so that a principle of causality is violated whenever an atom emits light, or a uranium nucleus decays, even though nothing that happened beforehand made this inevitable. Feynman (1967: 147) urged philosophers to acknowledge that this implication of quantum mechanics undermines the view that causal determinism forms a precondition of scientific inquiry. But while some physicists (notably Bohm 1957) have denied the implication, most philosophers since Reichenbach have accepted it with alacrity, and sought to develop accounts of causation equally applicable in an indeterministic or a deterministic world.

Such accounts typically appeal to probabilistic relations among events. But on closer examination, the way probability figures in quantum mechanics poses challenges to otherwise plausible philosophical accounts of causation.

These may be traced back to the famous EPR paper, in which Einstein, Podolsky, and Rosen (1935) argued against a prevailing interpretation of quantum mechanics as a theory that offers a complete description of an indeterministic world. EPR concluded that acknowledging the incompleteness of quantum mechanics's description freed one to seek a theory to complete this, and perhaps also restore determinism. This and similar arguments Einstein gave elsewhere rely on apparently plausible causal conditions intended to rule out what Einstein once referred to as 'spooky' action at a distance.

Later Bell showed how, by extending EPR's own line of argument, one could derive a contradiction between their locality assumptions and potentially testable predictions of quantum mechanics. Many of these predictions have subsequently been verified in ingenious experiments, so the theory of quantum mechanics may be 'factored out': the experimental results themselves apparently exclude any ordinary causal explanation of the sort that philosophers have focused on in their accounts of causation.

One response is to use a general account of causation to argue that causation in quantum mechanics has remarkable features—that cause and effect are not always connected by any continuous process, that quantum causes may act 'instantaneously' at a distance, or even that quantum causes may occur after their effects. A contrary response has been to add further conditions to a general account of causation, so the peculiar connections among distant events

involved in realizations of Bell–EPR-type situations are ruled non-causal. Others have drawn the more sceptical conclusion that it is a mistake to look to any general account of causation as a way of settling disputes about the nature and extent of causal relations in the quantum realm. Noting that our causal concepts were adapted to phenomena experienced far outside that realm, such sceptics argue that there is simply no uniquely correct way of applying these concepts in Bell–EPR-type situations, where their diverse components pull in different directions.

2. QUANTUM INDETERMINISM?

If quantum mechanics shows that the world is indeterministic, and an event is caused only if its past determined its occurrence, then quantum mechanics shows that there is little or no causation. But should we accept the antecedent? A few philosophers have endorsed its second conjunct, though these days most allow that an event may be caused even though it was merely rendered more likely by what came before. And the ready acceptance of the first conjunct by most philosophers of causation may be premature. There are reasons to question whether quantum mechanics establishes that our world is indeterministic.

Roughly, a theory is deterministic just in case its equations of motion have a unique solution for the state of any system at a later time, given the state at an earlier time. The Schrödinger equation—the basic equation of motion of the theory of non-relativistic quantum mechanics—has this feature. Margenau (1950), for one, concluded that quantum mechanics is deterministic! But the Schrödinger equation merely specifies how a vector or wave-function representing an undisturbed system's instantaneous state varies with time. The state vector ψ is only probabilistically related to values of measurable magnitudes by the *Born Rule*. In the simplest case this specifies that the probability of observing value q_i in a measurement of variable magnitude Q (such as energy, or distance from a fixed point) when a system is in a state represented by unit vector ψ is given by

$$Pr_\psi(Q = q_i) = |(\psi, \psi_i)|^2 \qquad (1)$$

Here ψ_i is a unit vector representing a state in which such a measurement would certainly yield value q_i, and $|(\psi, \psi_i)|^2$ is the square (modulus) of the projection of vector ψ_i onto ψ (it is a number between 0 and 1 inclusive, as required by its interpretation as a probability). Consequently, while the quantum state of any undisturbed system at one time uniquely determines its state at any other time, the state up to the time of a measurement (typically) fails uniquely to specify its outcome. Quantum mechanics is indeterministic in this more operationally relevant sense.

Moreover, according to von Neumann's notorious *projection postulate*, the effect of measurement in this simple case is discontinuously to alter the state vector from ψ to ψ_i, in contravention to the Schrödinger equation. But the disputed status of the projection postulate

within quantum mechanics is just one aspect of the outstanding problem of giving an adequate quantum mechanical account of the measurement process.

The history of physics includes many highly successful theories that later came to be superseded by theories with very different conceptual structures. The present empirical success of quantum mechanics is compatible with it eventually being subsumed as an approximately true limiting case of some even more successful deterministic theory.

Even if quantum mechanics were a true indeterministic theory, the world might still be deterministic if it contained additional parameters that, together with a system's quantum state, uniquely determined the outcome of a measurement on that system. Quantum orthodoxy maintains the completeness of the quantum state. But EPR challenged that claim, and Bohm (1952) formulated a theory in which the positions of particles acted as just such additional parameters, conventionally known as 'hidden variables' despite the fact that for Bohm these were the *only* variables that were ever directly measured. Bohm's deterministic theory is arguably empirically equivalent to, and perhaps offers an unconventional interpretation of, quantum mechanics itself. Physicists have objected to Bohm's theory: as we shall see, it exhibits a striking non-locality. But observation alone cannot definitively rule out the possibility of understanding quantum mechanics in a Bohmian way as a true theory of a deterministic world. Indeed, Bohm's is not the only way of so understanding it. One can take Everett's (1957) very different relative-state formulation of quantum mechanics as offering a theory of an objectively deterministic world that inevitably appears indeterministic to observers like us.

3. QUANTUM NON-LOCALITY?

Some quantum systems are composed of spatially dispersed subsystems. An experimenter can choose to measure any one of a variety of magnitudes on each subsystem at about the same time but at distant locations. These include linear polarization along independently selected axes of each photon in a pair emitted by a common source, as well as independently chosen spin-components of each of a pair of low-energy protons scattered off each other. Quantum mechanics predicts probabilities for different outcomes of any such measurement by applying the Born Rule to the state vector of the composite system: sometimes its subsystems fail to have state vectors of their own, in which case they are said to be 'entangled'. Many such probability distributions have been experimentally tested since 1972. Some such experiments have involved measurements on each of two photons separated by several kilometres: others (Aspect, Dalibard, and Roger 1982) have involved measurements performed on photons at space-like separation, so that not even light in a vacuum could travel from one event to the other.

By now the accumulated statistics provide overwhelming confirmation of quantum mechanical predictions. But if the separated subsystems are 'entangled' these statistics cannot always be explained in the usual way by appeal to a prior common cause unless there are direct causal connections between the separated subsystems, or at least between distant measurement events involving them. Since there is no evidence of any relevant mechanism

connecting these events, and no known theory (including quantum mechanics) describes one, some have claimed this demonstrates action at a distance: and since the events may be space-like separated, some have taken it as an example of superluminal causation, in apparent conflict with relativity theory. If they are right, then these non-localized statistical correlations exhibit quantum non-locality. While evaluating this thesis takes some work, this is effort well spent by philosophers interested in causation.

3.1 Bell Inequalities

Bell noticed a peculiar feature of Bohm's (1952) theory: it implied that the result of a measurement on one particle *here* sometimes depends on a measurement on another particle *there*, no matter how far apart or when (within limits) these measurements are made. In his classic paper (Bell 1964) he generalized this to a class of deterministic hidden variable theories, which supplement the quantum state by additional parameters sufficient uniquely to prescribe the result when any magnitude is measured on a quantum system. Bell (1971) and others subsequently further generalized his result to cover stochastic hidden variable theories, in which only the *probabilities* of measurement outcomes are determined by the suitably supplemented quantum state. In each case, the proof depended on certain *locality* conditions. He concluded that any hidden variable theory capable of reproducing the predictions of quantum mechanics after suitably averaging over the values of its additional parameters must, like Bohm's theory, be non-local.

Many related theorems have since been proved, to the effect that *no* theory meeting a number of independently plausible general conditions (some apparently having to do with locality) can reproduce all the predictions of quantum mechanics for the results of measurements of certain magnitudes on a variety of composite systems. It is therefore interesting to scrutinize each condition invoked by these proofs to see how far it may be justified by principles governing how causes operate. Three such conditions may be highlighted after explaining some notation.

Consider a pair of systems, at most one of which may be observed in each of two separate wings L, R of apparatus. The apparatus at each wing may be set independently to measure the value of no more than one of three different magnitudes labelled by $i, j = 1, 2, 3$, each with one of two possible outcomes labelled by $I, J = +, -$ for example, photon polarization along one of three different axes. So one possible result of a joint measurement would be $2 + 3$ -, where the setting and result of the left wing are written to the left of that at the right wing. Write the probability of outcome I in a measurement of magnitude i at the L wing and outcome J in a measurement of magnitude j at the R wing as $Pr_{ij}(I, J)$. The Born Rule gives the values of such probabilities, and experiment confirms them.

A stochastic model of the observed statistics would derive these probabilities by averaging over different possible theoretical states $\lambda \in \Lambda$ of the pair and perhaps other aspects of the experimental arrangement that are independent of what magnitudes the apparatus are set to measure. The probabilities are taken to arise as follows

$$Pr_{ij}(I, J) = \sum_{\lambda \in \Lambda} Pr_{ij\lambda}(I, J).Pr(\lambda) \qquad (2)$$

where $Pr_{ij\lambda}(I, J)$ is the probability the model assigns to outcomes I, J in measurements of i, j at wings L, R respectively in theoretical state λ. This implicitly assumes that the probability $Pr(\lambda)$ of state λ is independent of i, j: we shall examine this assumption later.

Further conditions on the theoretical probabilities may restrict the 'surface probabilities' $Pr_{ij}(I, J)$ derivable from the model. Two conditions are salient, and each has been motivated by causal considerations. First, define

$$Pr_{ij\lambda}(I, .) \equiv Pr_{ij\lambda}(I, +) + Pr_{ij\lambda}(I, -) \qquad (3a)$$
$$Pr_{ij\lambda}(., J) \equiv Pr_{ij\lambda}(+, J) + Pr_{ij\lambda}(-, J) \qquad (3b)$$

The first condition (or a close analogue)

$$Pr_{ij\lambda}(I, J) = Pr_{ij\lambda}(I, .) . Pr_{ij\lambda}(., J) \qquad \textit{(Outcome Independence)}$$

has also been called completeness, causality, and (confusingly!) factorizability in the literature. The second condition (or a close analogue)

$$Pr_{ij\lambda}(I, .) = Pr_{ij'\lambda}(I, .) \equiv Pr_\lambda(I)$$
$$Pr_{ij\lambda}(., J) = Pr_{i'j\lambda}(., J) \equiv Pr_{j\lambda}(J) \qquad \textit{(Parameter Independence)}$$

has also been called (hidden) locality. I have adopted Shimony's (1993) terminology, though his formulations differ in detail. Together, Outcome and Parameter Independence imply a third condition

$$Pr_{ij\lambda}(I, J) = Pr_{i\lambda}(I) . Pr_{j\lambda}(J) \qquad \textit{(Factorizability)}$$

Any surface probabilities derived from a stochastic model satisfying this Factorizability condition must meet a set of *Bell inequalities* (see e.g. Shimony 1993; 2004). But probabilities derived from the Born Rule do not always satisfy these constraints, and violations of Bell inequalities have now been amply confirmed.

3.2 Screening Off

Both Outcome Independence and Parameter Independence resemble screening-off conditions, where correlated (types of) events A, B are screened off by an event (of type) C if and only if

$$Pr\,(A\&B) \neq Pr\,(A).Pr\,(B) \qquad\qquad (Screening\ Off)$$
$$Pr\,(A\&B/C) = Pr\,(A/C).Pr\,(B/C)$$

Screening-off conditions have played a significant role in accounts of stochastic causation. They are central to Reichenbach's formulation of his common cause principle, which claims that if events of certain types are not directly causally related but nevertheless statistically correlated then they have a common cause—a distinct event of a type that (among other things) screens one off from the other. The observed violation of Bell inequalities has been held up as a striking counterexample to this claim.

Of course, it is a counterexample only if the statistical correlations are between types of distinct events that are not directly causally related. As we shall see, some have maintained that the relevant events are indeed directly causally related, even in cases where they are space-like separated—implying at least a prima facie conflict with relativity theory. Others have been tempted to deny that there are distinct events at the two wings to be correlated.

To make Outcome Independence look more like a screening-off condition, one could formulate it as follows:

$$Pr\,_{ij}(I\&J/\lambda) = Pr\,_{ij}(I/\lambda).Pr\,_{ij}(J/\lambda) \qquad\qquad (4)$$

This commits one to a larger probability space in which probabilities like $Pr_{ij}(\lambda)$ are defined, which is reasonable since the model assumed that $Pr(\lambda)$ is well defined and independent of i, j. But exactly what events are featured here? I and J are presumably localized happenings in short-lived but rather illdefined regions U_L, U_R of space-time overlapping the world-tubes of the left and right wings respectively. λ may be taken as the state the theory attributes to the pair as it leaves its source, or it may include additional information about events leading up to I and J, including events involving the apparatus at the two wings. Outcome independence fails if λ labels a generic quantum state vector of a pair. In the singlet spin state ψ_s, for example, we have $Pr_{ii\psi_s}(I, I) = 0$ but $Pr_{ii\psi_s}(I, .) = Pr_{ii\psi_s}(., I) = \frac{1}{2}$.

But this does not refute Reichenbach's principle unless the quantum state is the only candidate for a common cause here. We get a weaker version of Outcome Independence by taking λ as an enormous event that also includes everything that happens in or on the overlap of the past light-cones of U_L, U_R: certainly that would cover any Reichenbachian common cause of I and J compatible with a relativistic requirement that this absolutely precede its effects. If even this version of Outcome Independence were to fail, then that would surely constitute a significant counterexample to Reichenbach's principle.

Violation of Bell inequalities does not by itself refute Outcome Independence. In Bohm's theory, for example, the positions of the particles in the overlap of the past light-cones of U_L, U_R add relevant additional information to the quantum state, so that Outcome Independence holds (all the relevant probabilities are 0 or 1)! Bohm's theory is not constrained by Bell

inequalities, but only because it violates Parameter Independence. The general moral is that a violation of Bell inequalities yields a counterexample to Reichenbach's common cause principle only if it results from failure of Outcome Independence rather than some other condition required for their proof.

3.3 Act/Outcome Correlations?

Jarrett (1984) invoked relativity theory in defence of Parameter Independence. He argued that when the act of setting the apparatus and performing the measurement at the $R(L)$ wing is spacelike separated from the outcome in $U_L(U_R)$, then relativity prohibits any direct causal connection between these events. Nevertheless, in the experiments of Aspect, Dalibard, and Roger (1982) the predicted violations of Bell inequalities were observed in just these circumstances. But does relativity prohibit direct causal connections between spacelike separated events? And need any such connections be involved in violations of Parameter Independence?

A variety of proofs (e.g. Bell 1987: 60; Redhead 1987: 113–16) use quantum mechanics to show that no manipulations performed solely at one wing of an Aspect-type experiment are correlated with detectable variations in conditions at the other wing: there is no way of using the predicted correlations to set up a 'Bell telephone'. But this rules out direct causal connections only if these would permit act/outcome correlations of this kind.

In Bohm's (1952) theory, carrying out a measurement in the $R(L)$ wing will typically influence the subsequent trajectory of the system in the $L(R)$ wing. But this trajectory is made manifest only by the outcome at the $L(R)$ wing, which cannot be compared to what it would have been had a different measurement been carried out at the $R(L)$ wing. Because it violates Parameter Independence, there are act/outcome correlations between spacelike separated events in Bohm's theory, vindicating the usual view that it involves direct causal influences between these events. But does relativity rule out *all* models of Bell–EPR-type correlations that violate Parameter Independence?

There are two sorts of reasons to doubt that it does. Maudlin (1994) and others have questioned whether relativity alone prohibits causal connections between spacelike separated events. If relativity theory merely specifies the structure of spacetime and constrains the form of all physical laws, then its statement involves no causal notions and the theory can, by itself, have no causal implications. Causal relations between spacelike separated events would have certain peculiar features: cause and effect would be neither absolutely simulataneous nor have a frame-independent time order. But these generate a 'causal paradox' only together with additional assumptions, of which the argument could be treated as a *reductio*.

Even if relativity does rule out direct causal connections between spacelike separated events, violations of Parameter Independence need not require them. Again, one must add explicitly causal assumptions to extract this requirement from a violation of that condition. To clarify matters, recast Parameter Independence in the form of the following screening off conditions:

$$Pr\,(I\&j/i,\lambda) = Pr\,(I/i,\lambda).Pr\,(j/i,\lambda) \qquad (5a)$$
$$Pr\,(i\&J/j,\lambda) = Pr\,(J/j,\lambda).Pr\,(i/j,\lambda) \qquad (5b)$$

Here the probability space has been further enlarged to assign probabilities to events including the setting and performance of measurements. This may be challenged on the grounds that no probabilities should be assigned to events that are under the control of the experimenter. But suppose on the contrary that the setting at *L* and the outcome at *R* had some hidden, unknown stochastic common cause in the overlap of their past light-cones. Then one would not be surprised at the violation of the second of these last conditions, even if one did believe that relativity ruled out superluminal causation. It requires an additional causal assumption to rule out such a 'conspiracy theory'. Nevertheless, I expect readers to agree with Bell (1987: 100–3) that appealing to such a conspiracy theory to explain violations of Parameter Independence consistent both with alleged causal requirements of relativity and with Reichenbach's common cause principle is grasping at straws.

3.4 Outcome/Outcome Causation?

It is now widely acknowledged that measurement outcomes whose statistics violate Bell-inequalities manifest a peculiarly intimate connection among spatially separated 'entangled' quantum systems. But there is no consensus on whether this establishes a causal link between these measurement outcomes. This subsection considers arguments in favour of a causal connection: arguments against are evaluated in the next section.

Causal relations are intimately associated with counterfactual conditionals (see Ch. 8 above). Lewis (1973; 1986) offered an influential counterfactual analysis of causation. Butterfield (1992) argued that it is a consequence of this analysis that (if there is no act/outcome causation then) outcomes at different wings of a Bell–EPR-type situation are directly causally connected, even when spacelike separated. While himself rejecting such superluminal causation, he noted that a defender of Lewis's analysis might seek to avoid this consequence by denying that these outcomes are distinct events. But this is both implausible and difficult to reconcile with Lewis's own (1986) account of the nature of events.

Maudlin (1994: 130) defended a sufficient condition for causal connection of distinct events modelled on Bell's (1987: 54) *local causality*—that the state of its past light-cone should screen off event *A* from any spacelike separated event *B*. He argued that if there is no act/outcome causation then his condition is met by the perfectly correlated outcomes in some EPR-type situations.

Like Lewis's analysis, Maudlin's condition implies that certain spacelike separated outcomes at the two wings are causally connected—if not directly, then through a common cause that lies outside the past of at least one outcome. Maudlin applies his condition to a common understanding of quantum mechanics, in which definite measurement outcomes are secured not by hidden variables but through a stochastic process grounding the projection postulate. He concludes that, on this understanding, it is this stochastic process that provides the causal connection between the distant measurement outcomes, however hard it may be to square its occurrence with the demands of relativistic invariance.

4. No Causation?

Physicists have generally been reluctant to accept causal connections among space-like separated events, including measurement outcomes in Bell–EPR-type situations. Shimony (1993: 151) maintained that 'Quantum mechanics violates Outcome Independence, which cannot be used to send a superluminal message; and it conforms to Parameter Independence, a violation of which would permit superluminal communication. We may summarize by adapting a well-known political slogan: there is "peaceful coexistence" between quantum mechanics and relativity theory.' Elsewhere (ibid. 133) he says,

If [the outcomes] have space-like separation, then a causal connection between them would be an action at a distance. And even if one abstains from using a causal locution in this situation, the very fact that the correlation cannot be attributed to the complete state λ constitutes a kind of nonlocality (which perhaps is *sui generis* and appropriately named 'passion at a distance').

After applying the suggestively named microcausality condition embedded in relativistic quantum field theory—the condition that observables pertaining to space-like separated regions commute—in his proof that violation of Bell–inequalities does not permit superluminal signalling, Bell himself (1987: 61) comments that 'in this *human* sense relativistic quantum mechanics is local'. Right after, he admits that the key *local causality* condition required in the accompanying proof of the inequalities themselves 'may not embody *your* idea of local causality. You may feel that only the "human" version of the last section is sensible and may see some way to make it more precise.' Now there is an old argument, reproduced by Malament (1996: 5), that failure of the microcausality condition entails statistical act–outcome correlations that would permit superluminal signalling: so perhaps microcausality itself makes the 'human' version precise?

Bell notes that while relativistic quantum field theory is generally thought to be fully compatible with relativity, it does *not* meet his condition of *local causality*. Indeed, Bell–inequalities are now known to be violated by various states of a relativistic quantum field (see e.g. Clifton and Halvorson 2000). The desire to secure Shimony's 'peaceful coexistence' therefore provides a powerful motivation for a physicist to deny that violation of Bell's *local causality* condition is a sufficient condition for there to be a direct causal connection between space-like separated events.

Redhead (1987) claimed that for a stochastic relation between distinct events to be causal it must meet a condition he calls *robustness*. He argued that since this condition is not met by measurement outcomes in Bell–EPR-type situations, these events are not causally connected. Here is his condition:

A stochastic causal connection between two physical magnitudes *a* and *b* pertaining to two separated systems *A* and *B* is said to be *robust* if and only if there exists a class of sufficiently

small disturbances acting on *B*(*A*) such that *b*(*a*) screens off *a*(*b*) from these disturbances. (ibid. 102)

Now this is not a plausible necessary condition on *partial* causes, as Redhead (1989) effectively acknowledged when he considered 'restoring robustness by regarding [the quantum state of the Bell–EPR pair] as contributing cause of the correlation'. He there deployed a modified robustness condition in attempting to rule out even this possibility. But the attempt has been criticized by Maudlin (1994) and others. While Maudlin finds no robustness condition defensible, Healey (1992) offers a conditional defence of a robustness principle based on a Kantian conception of causal laws as unconditionally prescribing probabilities for events, while maintaining that Redhead's (1989) use of this principle was illegitimate.

Failing any direct causal connection between measurement outcomes in Bell–EPR-type situations, it remains tempting to attribute their correlations to an earlier common cause. But any common cause λ that satisfies Factorizability seems ruled out by observed violations of Bell inequalities. When such observations are said to refute the class of local realistic theories (e.g. by Shimony 2004), 'realism' is often understood to require a Factorizable common cause in the overlap of the past light-cones of the measurement events, while 'locality' excludes any direct causal connection between them.

Van Fraassen (1989) took the existence of a common cause satisfying Outcome Independence as explicating epistemic realism's requirement that such correlations be explained by a causal mechanism. He used violation of Bell inequalities to argue against epistemic realism, since the common cause pattern of explanation fails for certain quantum-mechanical phenomena. But Chang and Cartwright (1993) objected that a common cause need not screen off one of its effects from another if it is not connected to them by a causal process that is continuous in space and time—a condition, they claim, that quantum mechanical processes fail to satisfy. They concluded that failure of Outcome Independence is no reason to reject a common cause explanation of Bell-inequality-violating correlations.

5. CAUSAL SCEPTICISM?

Noting philosophers' continuing failure to agree on an analysis of causation, one may doubt we have any clear and univocal notion. Skyrms (1984) sketched seven philosophical theories of causation and pitted them against the quantum phenomena we have been considering. After winnowing these down to three—statistical causation, transfer of a conserved quantity such as energy-momentum, causation as manipulability/signalling—he concluded that

our ordinary, every day conception of causation is an amiably confused jumble of all three, with the principles of locality of causation and temporal priority of cause to effect and perhaps a few other things thrown in for good measure. ...

When we ask whether causes operate locally in the quantum domain, the old cluster concept loses its heuristic value and becomes positively misleading. ...

Metaphysics must take lessons from Mother Nature. (ibid. 284)

Healey (1992: 193) went further, arguing that

One can have a complete physical account of the correlations manifested in an Aspect-type experiment and still not know how to characterize the causal relations between [the event e_s of emission of a particle pair and the measurement events e_L, e_R], because of the 'open texture' of the concept of causation. And this shows that causation is not a physical relation.

He proposed a non-separable explanation of the correlations suggested by his 1989 'modal' interpretation of quantum mechanics, and concluded that 'if the nonseparable model of the correlations is correct, then e_s is indeed a common cause of e_L and e_R' (ibid. 200), but that 'no concept of causation is uniquely applicable to this case, and so even within the nonseparable model of the correlations manifested in Aspect-type experiments there is no determinate answer to the question as to whether or not e_L and e_R are directly causally related' (ibid. 203).

Healey (1994) amplified this conclusion, motivating two distinct conceptions of causal explanation and arguing that one measurement outcome figures in a causal explanation of another according to just one of these conceptions.

6. BACKWARD CAUSATION?

Reichenbach took for granted that any cause of simultaneous but separated events must precede them. But accounts of EPR–Bell correlations have been entertained involving causes that succeed their effects. Such appeal to 'advanced action' may avoid superluminal causal propagation only by the drastic expedient of allowing a cause to have effects in its backward light-cone! But some philosophers have argued that backward causation is indeed at least a conceptual possibility (see Huw Price and Brad Weslake, Ch. 20 above).

Price (1996) himself advocates a form of advanced action as offering the prospects of a novel kind of local realistic account of violations of Bell-inequalities. This rejects the assumption that the 'hidden variable' distribution $Pr(\lambda)$ of state λ is independent of the settings i, j, even when these are freely chosen after emission of a particle pair. He defends an advanced action model against philosophical objections—that it involves fatalism and/or the denial of free will, or that backward causation would inevitably lead to paradoxical 'causal loops'. But he admits that 'As it stands, quantum mechanics does not embody the backward influence which is required to avoid nonlocality' (ibid. 248).

FURTHER READING

Feynman (1967: ch. 5) gives a classic introduction to the use of probability in quantum mechanics. While Richard Feynman presents the case for quantum indeterminism, David

Bohm (1957) maintains that the case is not proven.

The edited volume Wheeler and Zurek (1983) contains reprints of many important papers on quantum non-locality, including Einstein, Podolsky, and Rosen (1935), Bell (1964), and Bohm (1952).

Most of John Bell's classic papers, including his (1964), (1971), (1975), and (1977), are reprinted in Bell (1987). Abner Shimony (1993) contains several key papers as well as a nice exposition of Bell's theorem: see also Shimony (2004). Cushing and McMullin (1989) contains many interesting papers by both physicists and philosophers. Michael Redhead's (1987) is a good introduction to quantum non-locality.

David Lewis's (1973) influential analysis of causation is reprinted with appendices in his (1986). The last two chapters of Huw Price's (1996) present the case for advanced action in quantum mechanics.

REFERENCES

ASPECT, A., DALIBARD, J., and ROGER, G. (1982). 'Experimental Test of Bell's Inequalities Using Time-Varying Analyzers', *Physical Review Letters* 49: 1804–7.

BELL, J. S. (1964). 'On the Einstein Podolsky Rosen Paradox', *Physics* 1: 195–200.

—— (1971). 'Introduction to the Hidden Variable Question', *Foundations of Quantum Mechanics*. New York: Academic Press.

—— (1975). 'The Theory of Local Beables', TH-2053-CERN, 28 July; repr. in *Epistemological Letters* 9 (1976).

—— (1977). 'Free Variables and Local Causality', *Epistemological Letters*, February.

—— (1987). *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press).

BOHM, D. (1952). 'A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables, I and II', *Physical Review* 85: 166–93.

—— (1957). *Causality and Chance in Modern Physics*. London: Routledge, Kegan Paul.

BUTTERFIELD, J. (1992). 'David Lewis Meets John Bell', *Philosophy of Science* 59: 26–43.

CHANG, H., and CARTWRIGHT, N. (1993). 'Causality and Realism in the EPR Experiment', *Erkenntnis* 38: 169–90.

CLIFTON, R., and HALVORSON, H. (2000). 'Generic Bell Correlation between Arbitrary Local Algebras in Quantum Field Theory', *Journal of Mathematical Physics* 41: 1711–17.

CUSHING, J., and MCMULLIN, E. (eds.) (1989). *Philosophical Consequences of Quantum Theory: Reflections on Bell's Theorem*. Notre Dame, Ind.: University of Notre Dame Press.

EINSTEIN, A., PODOLSKY, B., and ROSEN, N. (1935). 'Can Quantum-Mechanical Description of Physical Reality be Considered Complete?' *Physical Review* 47: 777–80.

EVERETT, H. W., III (1957). 'Relative State Formulation of Quantum Mechanics', *Reviews of Modern Physics* 29: 454–62.

FEYNMAN, RICHARD (1967). *The Character of Physical Law*. Cambridge, Mass.: MIT.

HEALEY, R. (1992). 'Chasing Quantum Causes: How Wild is the Goose?' *Philosophical*

Topics 20: 181–204.

—— (1994). 'Nonseparable Processes and Causal Explanation', *Studies in the History and Philosophy of Science* 25: 337–74.

JARRETT, J. (1984). 'On the Physical Significance of the Locality Conditions in the Bell Arguments', *Nous* 18: 569–89.

LEWIS, D. (1973). 'Causation', *Journal of Philosophy* 70: 556–67.

—— (1986). *Philosophical Papers II*. Oxford: Oxford University Press.

MALAMENT, D. (1996). 'In Defense of Dogma ...', in R. Clifton (ed.), *Perspectives on Quantum Reality*. Dordrecht: Kluwer Academic, 1–10.

MARGENAU, H. (1950). *The Nature of Physical Reality*. New York: McGraw-Hill.

MAUDLIN, T. (1994). *Quantum Non-Locality and Relativity*. Oxford: Blackwell.

PRICE, H. (1996). *Time's Arrow and Archimedes' Point*. Oxford: Oxford University Press.

REDHEAD, M. (1987). *Incompleteness, Nonlocality and Realism.* Oxford: Clarendon.

—— (1989). 'Nonfactorizability, Stochastic Causality and Passion-at-a-Distance', in Cushing and McMullin (1989: 145–53).

SHIMONY, A. (1993). *Search for a Naturalistic World View*. Cambridge: Cambridge University Press, ii.

—— (2004). 'Bell's Theorem', *Stanford Electronic Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/bell-theorem/>, accessed 20 March 2009.

SKYRMS, B. (1984). 'EPR: Lessons for Metaphysics', in P. A. French, T. E. Uehling, and H. K. Wettstein (eds.), *Midwest Studies in Philosophy*. Minneapolis: University of Minnesota Press, ix. 245–55.

VAN FRAASSEN, B. (1989). 'The Charybdis of Realism: Epistemological Implications of Bell's Theorem', in Cushing and McMullin (1989).

WHEELER, J., and ZUREK, W. (eds.) (1983). *Quantum Theory and Measurement*. Princeton: Princeton University Press).

CHAPTER 34
CAUSATION IN SPACETIME THEORIES

CARL HOEFER

1. INTRODUCTION

As far as we know, all cause–effect relations are to be found in space and time; nothing affects us from outside, and nothing inside gets out to affect anything else. So if this chapter were about 'Causation in Spacetime', it could swallow most, if not all, of the rest of this volume. But instead our topic is causation in spacetime *theories*; a much more restricted topic, though still one that could occupy several books.

Although Russell maintained that causation was not to be found in advanced physical theories (we will come back to this below), even he would have admitted that, if one must talk of cause–effect relations between events, then spacetime theories may well place *constraints* on what sorts of causal relations may exist and how they may be arranged in time. They may also imply the possibility of surprising and unexpected causal relations, and even serious causal anomalies. In this survey we will look at what the three most important spacetime theories imply about causation. We will start with a brief look at Newtonian physics, see how important changes are introduced by Special Relativity theory, and finally turn to the rich causal fields of General Relativity models.

2. NEWTONIAN PHYSICS

Since the idea of spacetime is a conceptual novelty of the twentieth century, it may seem odd that our discussion begins with classical mechanics and gravitation. For Newton, space ('absolute space') was one thing, time was quite another, and there certainly was no single entity composed of the two together. But in fact, shortly after the birth of Einstein's relativity theories, physicists began to recast Newtonian physics as a spacetime theory, and in doing so achieved important conceptual insights (see Friedman 1983). It is therefore perfectly apt to talk about Newtonian spacetime, and what sorts of causality it permits and/or forbids.

The details of how one makes a spacetime out of Newton's space and time, and the difference between 'full' Newtonian spacetime and 'Neo-Newtonian' (or 'Galilean') spacetime, are not pertinent to our discussion. So we will not explain the structure of these spacetimes in detail, and instead will focus on the features of Newtonian physics that make for interesting causal facts. Chief among these features are the *absoluteness of simultaneity* and the *instantaneous action at a distance* permitted by the spacetime structure and assumed in the

law of gravitation.

In and , we see a depiction of the simultaneity structure of Newtonian spacetime, and the instantaneous actions-at-a-distance permitted by that structure (and posited through Newton's law of gravitation).[1] The gravitational force F_{ab} between bodies a and b, at a certain moment in time (say, midnight of 1 January 1999) is a function of the masses of a and b at exactly that time, and the distance between them at exactly that time. In we see a smooth evolution of the relative distances and gravitational forces between a and b; but if we could intervene, grabbing a and wiggling it back and forth, there would be an *instantaneous* difference in the force exerted by a on b, and hence an instantaneous change in the motion of b from what it would have been without the wiggling of a.
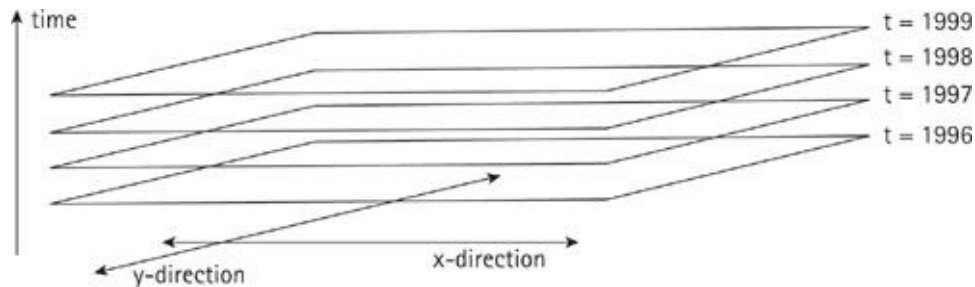
Fig. 34.1a Simultaneity surfaces of Newtonian Spacetime: Each slice represents all of space at one moment of time.
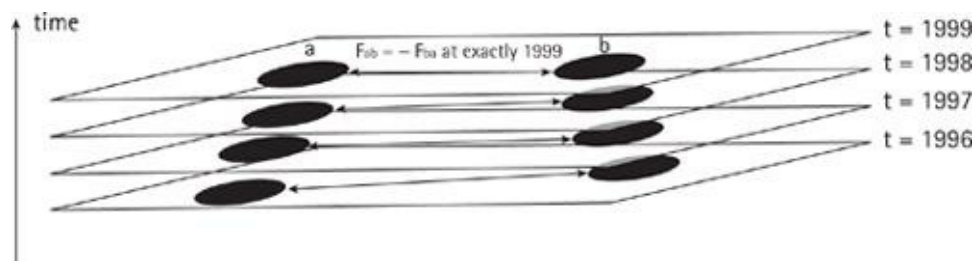
Fig. 34.1b Gravity force is 'instantaneous', determined by separation on the time-slice.

This instantaneous action caused more consternation in the Age of Enlightenment because of its at-a-distance or unmediated character, than for its instantaneous character. There was no reason at that time to suppose that an event at location p could not instantly affect events at distant location q—as long as there was a material connection between p and q. Grab a long stick and wiggle it *here*, and you can instantly (as far as anyone knew at the time) move the other end of the stick, poking someone *there*.

Today the intuitions of physicists are exactly the opposite. We are used to all interactions between particles being mediated by *fields* of one sort or another, and Newtonian gravitation lends itself to such treatment quite easily. On the other hand, as we will see, Einstein's relativity theories appear to rule out any instantaneous transmission of matter or energy across space, such as was implied by the description of the rod-wiggling case. Whereas Newtonian

spacetimes impose no upper limit on the speed at which a causal intervention may 'propagate' to another place, relativity theory sets an apparently strict speed limit: the velocity of light, c = 300,000 km/sec.

The lack of a speed limit in classical physics, coupled with the infinite/unbounded nature of space, makes possible certain causal anomalies—violations of determinism, in many cases—that have puzzled and either delighted or infuriated philosophers of physics in recent decades (depending on how strongly they desired classical physics to be a domain in which determinism reigns).[2] We will briefly look at three of these anomalies.

2.1 Space Deserters/Invaders

As Marc Lange discusses in Ch. 31 above ('Causation in Classical Mechanics'), Newtonian physics admits solutions in which a set of point particles moving under gravitational attraction alone can accelerate each other to unboundedly high speeds in a finite time—at the endpoint of which they have all effectively disappeared from the universe. Given that the time-reverse of any solution to Newton's laws is also a solution, this means that 'space invaders' could at any moment spontaneously appear in a Newtonian spacetime, zooming in 'from infinity' and, quite clearly, violating determinism.

2.2 Indeterminateness of the Infinite-Body Scenario

With spacetime theories come cosmologies, and Newton's physics brought into view the first recognizably 'modern' cosmologies, in which no special *centre* is postulated. Naturally, one of the first large-scale configurations of matter to be entertained was a homogeneous (at large scale) distribution of an infinite collection of bodies spread out over all space (see Fig. 34.2). But gravitational force is *universal* and cannot be screened or shielded off, and though it diminishes over distance as $1/r^2$, its reach is nevertheless infinite. Therefore the net gravitational force acting on every body should be the sum of the infinity of component forces exerted on it by all other bodies. But now a familiar fact from the mathematics of infinite sums bites us: under certain conditions, the sum is not well defined, but instead depends on the order in which the terms are added. The case of Newtonian forces from a homogeneous distribution of bodies is precisely one such case. Once again the peaceful causal order of classical physics suffers a breakdown due to the causal possibilities opened up by Newtonian spacetime's structure—here, its infinite, unbounded, and Euclidean spatial character.

The breakdown may be viewed as only apparent, if one is willing to translate the theory into a modern differential geometry-based version; but it is not clear then whether one should think of the theory as literally the same theory or not.[3] In any case, infinities of a different sort, if we allow them, engender still further breakdowns of causal order.

2.3 Infinite Particle Collections and Supertasks

A 'supertask' is a task that involves an infinite series of discrete component parts, but

which is performed in a finite span of time. Supertasks of various types have been a source of paradox and puzzlement for philosophers at least since the time of Zeno, but large-scale production of models of physical theories involving supertasks began in the 1990s and continues today, particularly in the work of J. P. Laraudogoitia. As Lange discusses (Ch. 31 above), Laraudogoitia (1996) considered an infinite collection of identical point particles arranged Zeno-style in a line along the x-axis: a particle at $x = 1$, another at $x = \frac{1}{2}$, the third at $x = \frac{1}{4}$, and so forth (see Fig. 34.3 below). At $t = 0$, they are all at rest. Another identical particle approaches this collection from the right with velocity v (situation A), striking the rightmost particle and initiating a chain of billiard-ball-style collisions, each collision transmitting all momentum to the particle immediately to the left and leaving the incoming particle at rest. Under one way of modelling particle collision, after v seconds the result is this: an infinite collection of particles, all at rest and distributed Zeno-style in the space $x = 0$ to $x = 1$ (situation B).
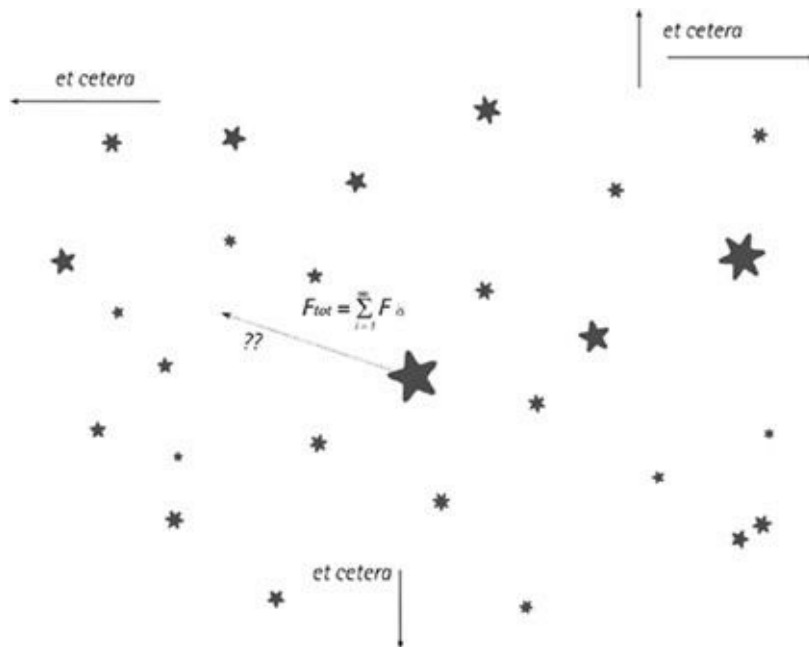
Fig. 34.2 With infinite bodies, what is the net gravitational force?

The temporal inverse of this supertask then consists of an infinite collection of particles distributed Zeno-style at rest in the unit interval (again, situation B), just sitting there for an arbitrary period of time, and then spontaneously 'self-exciting', starting to collide with one another and ultimately ejecting a particle off to the right (situation C). The violation of determinism is clear. In this case the element of spacetime structure that permits the causal anomaly is not spatial infinity or infinite velocity, but rather spacetime's *infinite divisibility* (which comes from the assumed continuity of the underlying manifold). If space possessed smallest bits or grains, then this scenario (and most other supertasks) could not be formulated. (It is equally true, of course, that if particles cannot be arbitrarily small, the supertask is not possible.)
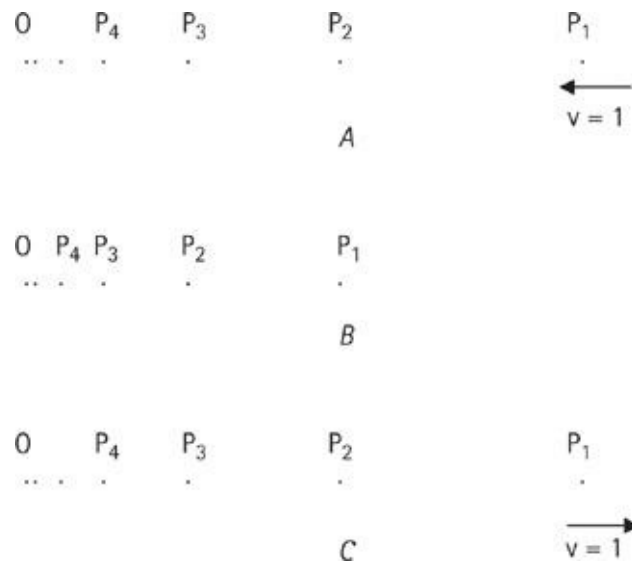
Fig. 34.3 A supertask and its time-inverse.

The three causal anomalies we have just considered are connected with diverse features of Newtonian spacetime: its lack of a speed limit for causal processes (which is, as we will see, nearly the same thing as its absolute simultaneity structure); its spatial infinity; and its infinite divisibility. In the move from Newtonian spacetime to Special Relativistic spacetime, the first of these features is removed.

3. SPECIAL RELATIVISTIC (MINKOWSKI) SPACETIME

Earman (1986; 2007) presents special relativistic physics as a relative paradise for determinism.

For the symmetries of Minkowski spacetime are given by the Poincaré group, which admits a finite invariant speed c, the speed of light, making it possible to formulate laws of motion ... which propagate energy-momentum no faster than c. For such laws all of the threats to classical determinism that derive from unbounded velocities are swept away. (2007: 29)[4]
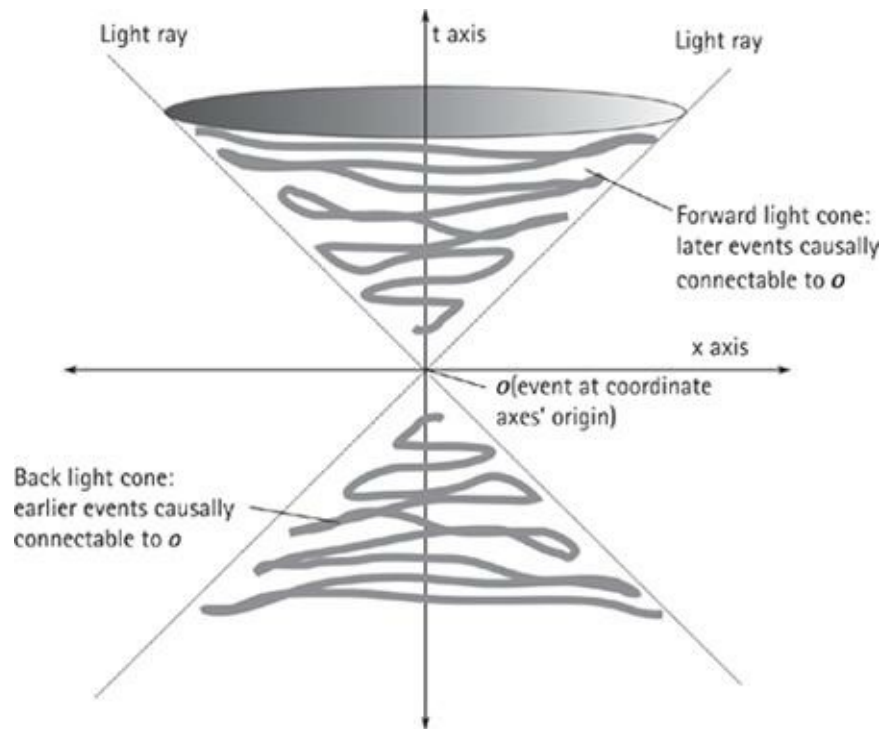
Fig. 34.4a Here we see a basic Minkowski diagram, illustrating how the paths of light-rays that travel 'in' to O or radiate 'out' from *O* determine the forward and back light cone structure at *O*. The forward light cone contains (the locations of) all events that *can* or *could* be causally affected by an event at O; the back light cone contains all events that *can* or *could* affect what happens at O. (The unshaded regions to the left and right are sometimes called the 'absolute elsewhere' of *O*; if FSP holds, no events there may affect or be affected by *O*.)

The postulate that no causal 'signal' may propagate through spacetime faster than c has come to be called the First Signal Principle (FSP). But it is important to realize that FSP does not follow logically from any combination of the following: the Relativity Principle, the Light Principle (saying that light rays propagate at c in any inertial reference frame), or the Principle of Special Relativity (saying that all physical laws should be Lorentz-covariant, and thus 'adapted' to the structure of Minkowski spacetime). Further substantive physical assumptions must be made in order to rule out causal connections with apparent velocities greater than c.
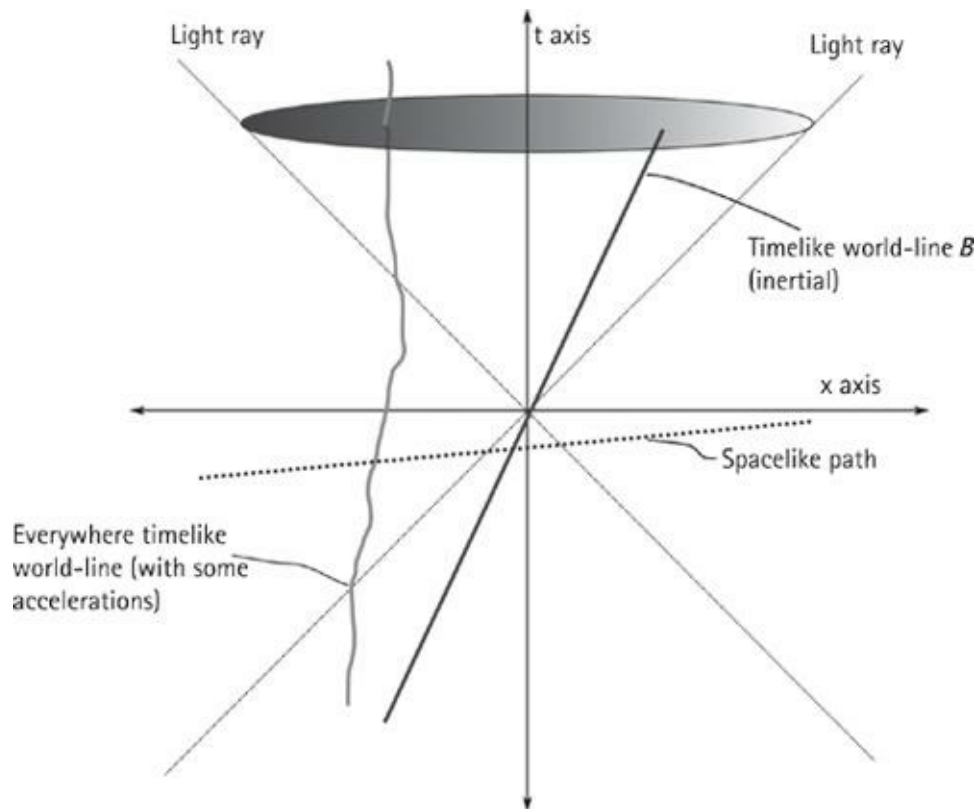
Fig. 34.4b Here we see examples of everywhere timelike paths, which represent physically possible world-lines for material objects, and one spacelike path, which represents the sort of connection ruled out by FSP.

We will assume for the rest of this section that all physics respects the FSP. This means that we may take the light-cone structure of Minkowski spacetime as equally representing the *causal* structure of spacetime, as illustrated in Figs. 34.4a–c.

The tight links between causality and spacetime structure normally postulated in discussions of SR inspired at least two prominent philosophical programmes that sought to use the links to make philosophical progress on longstanding conceptual problems. The first consists of the attempts, by Reichenbach (1956; 1958), Grünbaum (1968), and van Fraassen (1970) to mount a 'causal theory of time', providing a relationist-style reduction of temporal-relation facts to causal-relation facts.
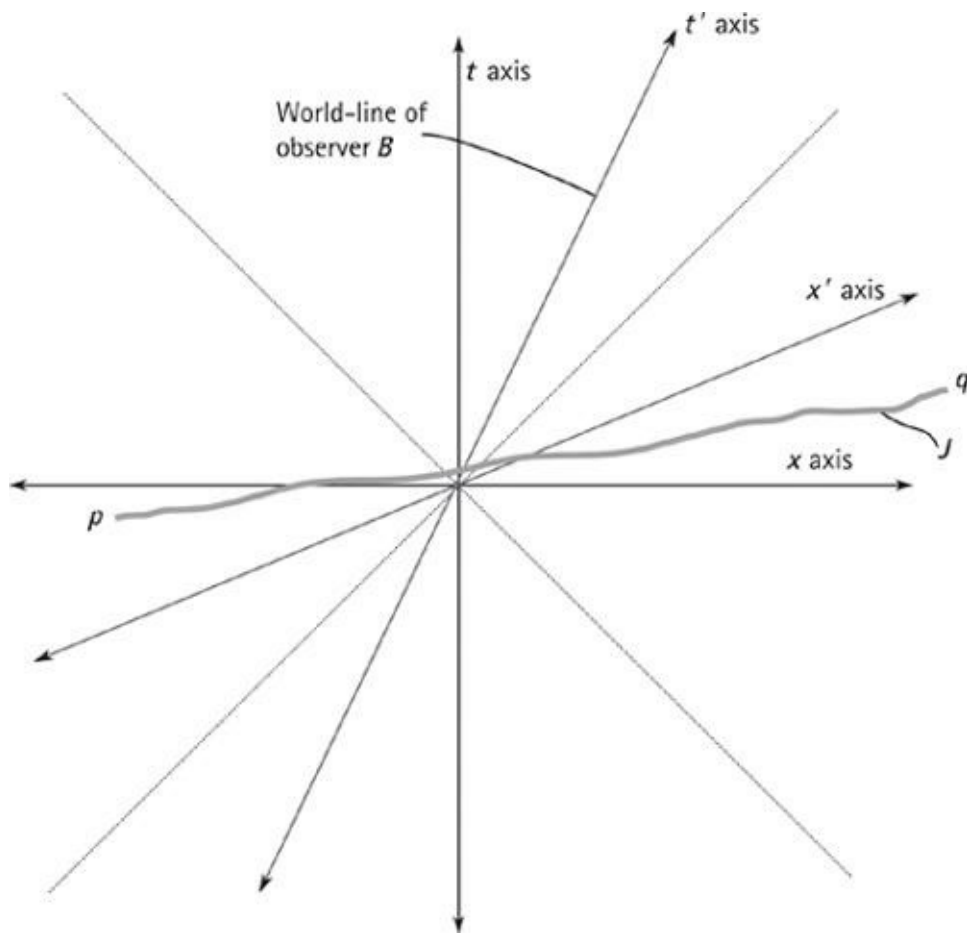
Fig. 34.4c Illustrates the relativity of simultaneity in SR: the t' axis is the path of an object (observer) moving inertially at high velocity relative to the x-t frame; the x'-axis illustrates a plane of simultaneity for the moving observer. The spacelike path J represents something moving, faster than light, from p to q—to an observer at rest on the x-axis. But to an observer moving with the $x' - t'$ frame, this path is either moving backwards in time from p to q, or forward in time (still faster than light) from q to p!

A causal theory of time seeks to reduce temporal concepts and relations to causal concepts and relations; the intuitive starting point is the observation, which can be seen already in Leibniz, that B is later than A if A is the cause of B. SR lent aid and comfort to this relationist perspective on time, for reasons we see clearly illustrated in Figs. 34.4a–c. At any point, events in the forward light cone of O and hence (absolutely) *later* than O, can be defined as the set of events possibly- causally-influenced by O; the back light cone (events *earlier* than O) is the set of events that possibly-causally-influence O; and events not possibly connected to O are spacelike separated from it, hence of *indeterminate* time order with respect to O.
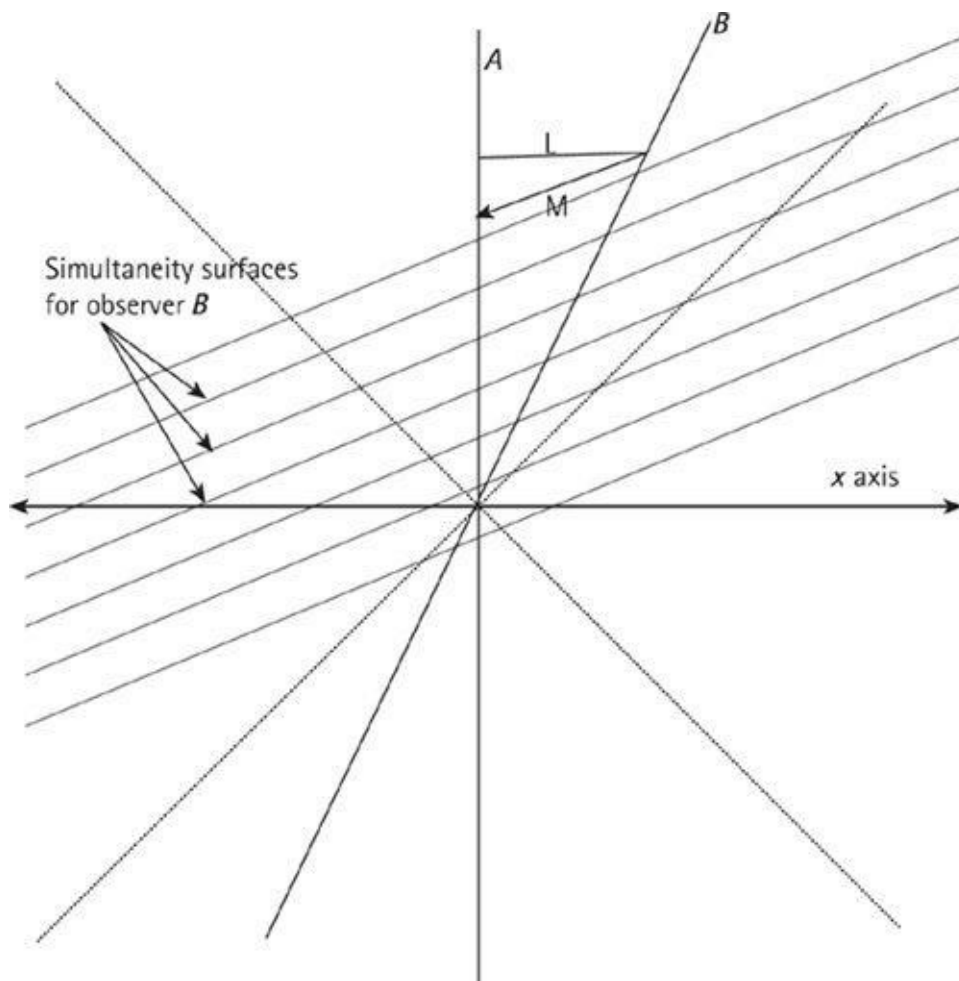
Fig. 34.5 Finally, we see one prominent reason why physicists prefer to rule out faster than light causal signals. By combining *L* (emitted by the rest observer *A*) with *M* (emitted by the moving observer *B*), where both *L* and *M* travel nearly-instantaneously (in the respective frames of their emitters), we can concoct scenarios of backwards causation, time travel, and the attendant paradoxes. Why can't *A* send a signal, passed on by *B*, telling his earlier self not to send that very signal? Or send a magic bullet back in time, to kill himself before he can send the bullet?[5]

Causal theories of time were popular during the heyday and waning years of logical empiricism, but came under increasingly severe attack in the late 1960s and 1970s (e.g. in Lacey 1968 and Earman 1972). While the theories' reductions were extensionally adequate in the context of the simple Minkowski spacetime structure of SR (the reduction was in essence first carried out by A. A. Robb (1914)), extension to the more complicated spacetime structures found in General Relativity (GR) was not clearly workable.[6]

The second philosophical programme closely connected to SR is the Salmon–Dowe programme of analysing causation in terms of *causal processes,* which in turn are analysed in terms of world-lines that carry a physically conserved quantity (see Salmon 1984 and Dowe, Ch. 10 above). Regardless of the success of this line of theorizing qua analysis of causation, it is of interest to us as another example of the tight connections between spacetime structure

and causation, for as an account of causation *in physics* the Salmon–Dowe approach is quite a natural one. In fact, it is again quite close to what we discussed in [Figs. 34.4a–c](#). But instead of offering an analysis of either time order or causal connectability, we can see Salmon and Dowe as offering an analysis of *actual* causal connection: events C and E are causally connected *iff* there is a chain of actual causal processes (world-lines of the right sort being *causal processes* and causal *interactions* defining the points where processes end/begin) connecting C to E. Again, the theory both takes inspiration from, and works well in, the context of SR; again, we will see below that things are not so clear-cut in the more complex arena of GR.

4. Spacetime Structures and Causation in GR

SR is not so much a physical theory as a set of principles, postulates, and demands, some of which we reviewed above.[7] Once the basic lesson of (local) Minkowski spacetime structure has been learned, the real pay-off comes in developing new physics to replace classical theories in the new setting. One branch of that effort leads in the direction of quantum mechanics and quantum field theory (see Teller 1995; Healey, Ch. 33 above). Here we will follow the other branch, that adapts gravitation to relativistic spacetime: General Relativity.

GR changes the nature of spacetime in two interrelated ways. First, the geometric structure of spacetime, as represented by the metric field tensor g^{ab}, is no longer a fixed 'background' in which events take place, but rather a variable, dynamical structure connected with the material contents of spacetime *via* Einstein's field equations. This leads to the second change: there are a huge variety of new cosmological model spacetimes, often with strange and exciting topological features, in which new causal puzzles and anomalies can arise. These two changes provide (in reverse order) the two themes we will discuss: in sect. 4.1 we will briefly look at some of the causal features of certain GR spacetimes; and in sect. 4.2 we will entertain the question of whether in GR, for the first time, ordinary matter and spacetime genuinely *causally interact* with one another.

4.1 Strange (Space)times

There is a huge variety of GR models that would merit consideration in this section, if space permitted. In fact, Earman's (1995) can be considered a whole book devoted to discussing the vagaries of causation and determinism in the overall context of permitted GR model worlds. We will have to touch just briefly on some key points.

First, we should note what a great difference there is between SR and GR. As noted above, unlike SR, GR is an actual physical theory, a theory of fluid dynamics and gravitation (to which electromagnetism can easily be added). There are various principles and postulates that played heuristic roles in its discovery, but every one of them has at most a dubious status in the finished product.[8] And the basic postulates or principles from SR vanish from sight. The Light Principle is, typically, only approximately true, in part due to the fact that inertial

frames are only well-defined 'locally', that is, over small-to-infinitesimal regions of spacetime (depending on local curvature). As Brown (2005: ch. 9) points out, it is not even universally true that light (e-m radiation) propagates along the 'light cones' as determined by the metric. The Special Principle of Relativity is replaced by the Weak Equivalence Principle (which enjoys solid empirical confirmation, but which could fail to be universally true without this undermining GR as a whole). The First Signal Principle seems secure, thanks in part to the same substantive assumptions that were needed to guarantee it for SR—but only if causal near-loops are prohibited, and these do occur in some models of GR.

Notoriously, Gödel's (1949) model cosmology has everywhere timelike paths that go 'all the way around spacetime' and return to their starting point. Though the picture is not faithful to the actual structure of a Gödel world, Fig. 34.6a shows how one can conceive of this: time can have the topological structure of a circle rather than an infinite line.[9] In 34.6b we see how, even if such timelike paths cannot quite get back to where they started (as is the case in other GR models), they permit clear violations of the FSP. Observer A can effectively 'send a signal' from C to E along the spacelike dashed line, by sending the signal along the timelike trajectory shown.
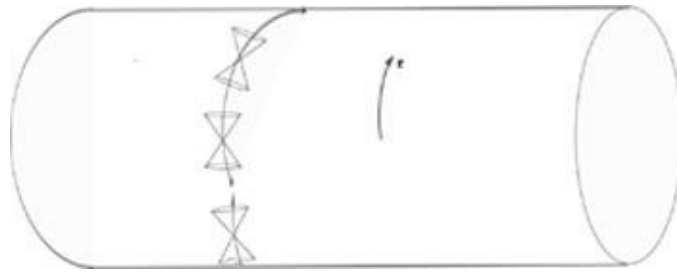
Fig. 34.6a A 'rolled-up' spacetime allowing everywhere timelike causal loops.
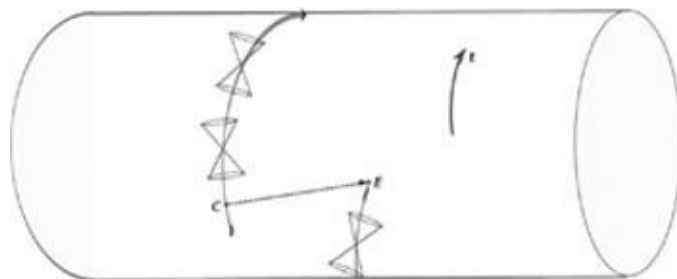
Fig. 34.6b Effective spacelike causal signalling, even though timelike loops are not possible.

While GR admits models with closed and near-closed timelike trajectories, those models all have features that rule them out as representations of our universe, or of any part of it. Does this mean that our universe is sure to be causally unproblematic? Hardly. Even if you do not find the Big Bang singularity (or Big Crunch, if our world turns out to have one) causally problematic, there are odd causal disconnects that are almost sure to be genuine features of our world. Consider two bodies moving inertially and initially separated by a finite distance. If

time is unbounded to the future, intuition suggests that if they last long enough, no matter how far apart they started, they will eventually 'see each other', that is, be causally connectable by a light ray. This expectation is refuted in some models with 'cosmological horizons': the expansion of spacetime is rapid enough that even light rays travelling indefinitely from one body toward the other, fail to arrive. (See Fig. 34.7a.) Moreover, if our world features any black holes (as astrophysical observations seem to suggest), then there are regions of spacetime—beyond the 'event horizon'—causally disconnected from the rest of the world in the sense that no lightlike or timelike possible trajectory can escape the hole: the singularity 'tilts the light cones' inward (Fig. 34.7b). Finally, although physicists would like to rule them out, and can do so with the help of extra postulates, Einstein's equations permit the existence of *naked* singularities—ones not hidden behind an event horizon—out of which in principle anything at all might emerge, ruining any hope of determinism. (See Earman (1995) for extensive, at times technically forbidding, discussion of all these matters.)
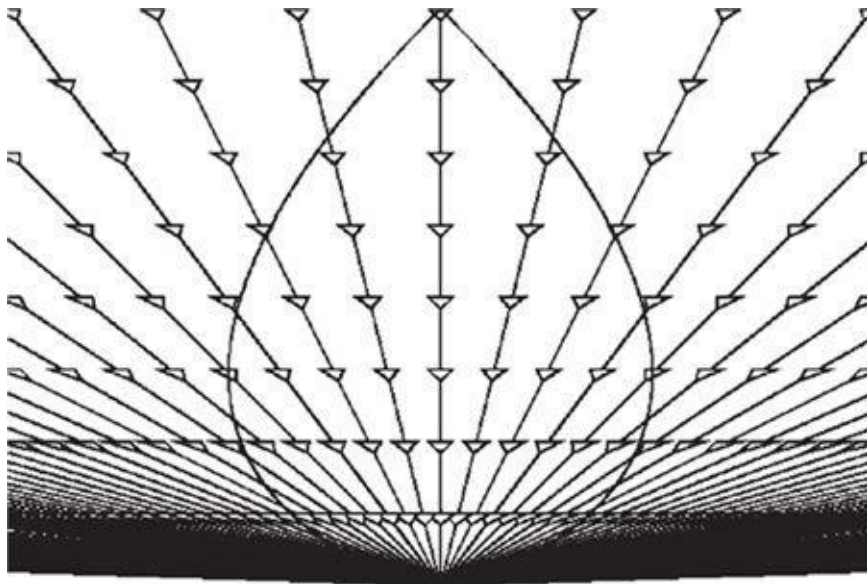
Fig. 34.7a Much of the universe is outside our back light cone; far enough away, in an expanding universe, galaxies exist that will *never* be inside our back light cone. (Image reprinted with permission, courtesy of Ned Wright: http://www.astro.ucla.edu/~wright/cosmo_03.htm>.)
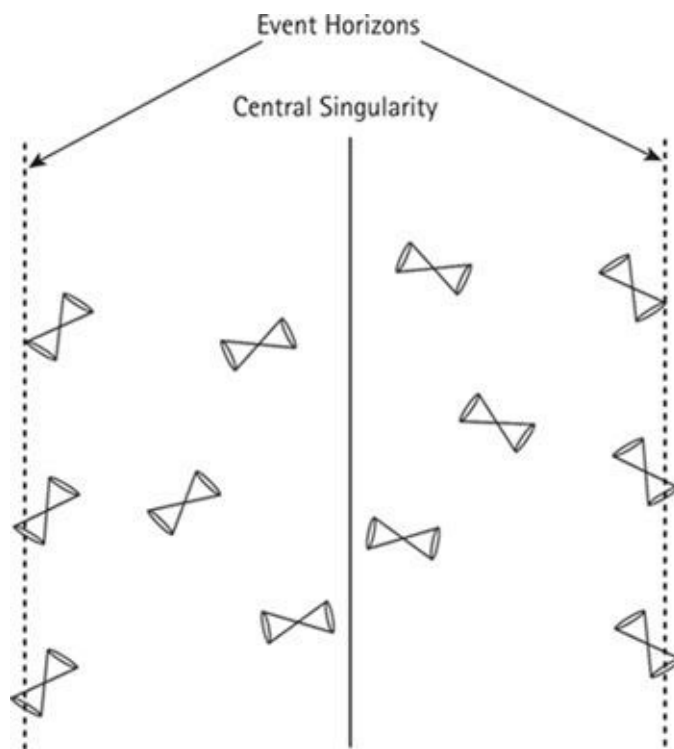
Fig. 34.7b Light cones are 'tipped' towards the black hole singularity; no future-directed timelike curve can cross the event horizon.

4.2 Matter–Spacetime Causal Interaction?

'Space acts on matter, telling it how to move. In turn, matter reacts back on space, telling it how to curve' (Misner, Thorne, and Wheeler 1973: 5).

In the context of pre-GR spacetime theories, most philosophers of science will allow that spacetime structure in some sense *explains* (or contributes essentially to an explanation of) the inertial motions of bodies not subject to external forces. It is far more controversial to cast this explanatory relation in *causal* terms; even such an arch-substantivalist as Graham Nerlich (1994) baulks at applying this term to inertial motion.

Einstein, however, did not mind this way of speaking; he insisted that in either Newtonian or SR physics, a *cause* of the distinction between inertial and non-inertial motion must be posited, and viewed the aloof and unobservable character of both Newtonian and Minkowski spacetimes as epistemologically unacceptable. Einstein therefore was pleased that his second spacetime theory made spacetime (again, in the form of the metric field tensor g^{ab}) epistemologically respectable by letting it both act *and be acted upon* by matter.[10]

Despite the common acceptance of this view of the matter-spacetime relationship in GR, there are reasons to regard this view as questionable, some of which we will briefly explore here.

As Russell (1913) argued, advanced physical theories—and GR is a paramount example— do not actually use the word 'cause' in any fundamental sense. (See Lange, Ch. 31 above, for further discussion of Russell's arguments.) SR can, perhaps be given an axiomatic

presentation in terms of possible causal connections, but GR certainly cannot, and Einstein's equations, while specifying the *mathematical* relationship between matter (T^{ab}) and metric g^{ab}, do not (of course) use the word 'cause', nor specify in which direction(s), if any, the cause–effect relationships go. Relativists use 'causal' in discussions of determinism, horizons, singularities, and so forth, but the term is easily eliminable from those discussions without substantive loss. These facts by no means show what Russell aimed to prove, namely that there is no causality to be found in advanced theories; but they will be worth keeping in mind for later.

There are three reasons to be uneasy about positing matter–spacetime causal interaction in GR. The first has to do with energy conservation. Recall the Salmon–Dowe analysis of causal interaction: exchange of a conserved quantity between causal processes. While allowing that the theory is not a plausible candidate analysis of causality *tout court*, we noted that it does seem to capture the intuitive notion of causal interactions in physics. The trouble, in GR, is that there is no genuine conservation of energy (see Hoefer 2000), but energy (momentum) is the only possibly conserved quantity that spacetime might possess and exchange with material stuff. In all realistic GR spacetimes, there is no *invariant* and well-defined quantity of gravitational energy-momentum that can be assigned to a spacetime point or finite region; the pseudotensors constructed to represent such energy, and the integrals based on them, are irremediably coordinate-dependent, which (according to relativists, when speaking of other things) is just to say, *not fully real*. When a binary star-pair spiral in towards each other and lose (genuine) energy-momentum, the standard presentation of what happens is that this energy-momentum is transformed into gravitational energy carried by g^{ab}, and carried off in gravitational waves. Later, other systems such as detectors may be affected by the waves, gaining back some of that energy. No harm would seem to be done by this way of speaking, but it should be recognized that the mathematics of the theory equally endorses a different gloss: The spiralling binary pair loses energy (full stop). Gravitational waves radiate outwards from the pair. Later/elsewhere, wave detectors may gain energy (full stop; the energy need not have 'come from' the gravity waves). Since GR has no genuine, integral-form conservation law, this gloss is no less legitimate than the former.

Relativists will generally prefer the 'exchange' gloss, since they spend great efforts trying to calculate the *amounts* of energy lost by binary pairs and carried off by gravity waves, and the amount of energy *gain* that should be expected at distant detectors. Idealizing fairly radically, they can calculate reasonably robust results that are 'invariant enough' in the right sort of spacetime (usually, one that is flat at infinity and basically only contains the binary pair). This should not satisfy philosophers; after all, quantum physicists predict results for measurements with great success, but (almost) nobody takes this as demonstrating that there is no measurement problem in QM.

The second sort of reason to question matter–spacetime causal interaction arises when we think of causation as intimately connected with counterfactuals (see Paul, Ch. 8 above). Taking a straightforward approach to reading counterfactuals off Einstein's equations, we get causal influence going in both directions just as Misner, Thorne, and Wheeler (1973) describe. 'Had there been twice as much matter here, the local curvature would have been twice as great.' 'If space were more curved here, free falling bodies would approach each other more rapidly'—and so forth. The trouble is that the equations of the theory equally support

counterfactuals whose causal correlates we would prefer not to accept. For example (and this was one of Russell's main points), the fact that GR permits retrodiction as much as prediction means that 'backtracking' counterfactuals are as commonplace as forward-looking counterfactuals, but we want to embrace only forward causation.[11] And when we set up typical what-if-things-here-now-had-been-different counterfactuals using the initial-value formulation of GR (the most natural way to proceed), the so-called *constraint equations* — which must be satisfied if the theory is to be used at all—entail that changes in spacetime and/or matter *here-now* not only entail differences *here-later*, but also differences *there-now*, where 'there-now' may be a region spacelike separated from *here-now*. But we do not believe, and will not accept, that things here causally influence those distant states of affairs. So again, we will have some work to do to purge the unwanted causality.[12]

Thirdly and finally, we should be cautious about embracing too literally the matter–spacetime causal interaction story because, at the end of the day, GR is a *phenomenological* theory rather than a fundamental theory. On the matter side of the field equations, T^{ab} is a fluid-dynamics representation of matter: continuous, non-quantum mechanical. We know that matter just isn't really like that, at a fundamental level. As for the other side of the equations, g^{ab} may not turn out to be a fundamental object either. Arguably, most efforts to quantize gravity presuppose that it is not. Which brings us to our final question: what may future theories teach us about causation in spacetime?

5. Conclusion

As philosophers of physics, we would like to understand not only *how* physics manages to represent the world with such empirical success, but also *what* the world must be like, intrinsically, given that these theories are successful. Not having any 'final physics' in hand is a serious obstacle for this effort. Certainly, our best theories are very successful and we ought to learn from them; but they are also, as Feyerabend stressed, *already refuted*. This goes as much for GR as for SR, as much for ordinary QM as for Newtonian mechanics. What good is it, then, to study the implications for causation of a refuted theory?

Well, the thing is not altogether desperate. The causal limitations first imposed by the speed limit c in SR survive basically intact in GR, and most physicists believe that they will survive in any quantum gravity successor to GR as well. Astrophysical observations already provide firm enough grounding for such things as black holes, cosmic expansion, and particle horizons, that we may be confident that they (or very near cousins) will survive into future physics as well. There will be no going back to the simple Euclidean world entertained prior to 1915; and that is a cause for satisfaction. Future theories will presumably bring other novel, permanent additions to our understanding of the relations between spacetime structure and causation.

Further Reading

The classic text to introduce all the spacetime theories discussed here remains Friedman

(1983). A quite different perspective on relativity theories, spacetime structure, relativity and conventionality of simultaneity, and the nature of the GR metric field can all be found in Brown (2005). For in-depth discussion of determinism and its fortunes in the main historically important physical theories, see Earman (1986) and the update/survey paper (2007). Causal anomalies and problems of all kinds in GR are most thoroughly discussed in Earman (1995). Finally, many of the topics touched on in this article are discussed with great clarity and rigour by Norton (2007).

References

BROWN, H. (2005). *Physical Relativity*. Oxford: Oxford University Press.

EARMAN, J. (1972). 'Notes on the Causal Theory of Time', *Synthese* 24: 74–86.

—— (1986). *A Primer on Determinism.* University of Western Ontario Series in the Philosophy of Science 32. Dordrecht: Reidel.

—— (1995). *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes.* Oxford: Oxford University Press.

—— (2007). 'Aspects of Determinism in Modern Physics' in J. Butterfield and J. Earman (eds.), *Philosophy of Physics*. Handbook of the Philosophy of Science Series. Amsterdam: Elsevier.

EINSTEIN, A. (1905). 'Zur Elektrodynamik bewegter Körper' ('On the Electrodynamics of Moving Bodies'), *Annalen der Physik* 17: 891–921.

FRIEDMAN, M. (1983). *Foundations of Space-Time Theories*. Princeton: Princeton University Press.

GÖDEL, K. (1949). 'An Example of a New Type of Cosmological Solutions of Einstein's Field Equations of Gravitation', *Reviews of Modern Physics* 21: 447–50.

GRÜNBAUM, A. (1968). *Modern Science and Zeno's Paradoxes*. Middletown, Conn.: Wesleyan University Press.

HOEFER, C. (1994). 'Einstein's Struggle for a Machian Gravitation Theory', *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 25: 287–336.

—— (2000). 'Energy Conservation in GTR', *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 31: 187–99.

—— (2003). 'Causal Determinism', Stanford Encyclopedia of Philosophy (online) http://plato.stanford.edu/entries/determinism-causal/>, accessed 24 April 2009.

LACEY, H. M. (1968). 'The Causal Theory of Time, a Critique of Grünbaum's Version', *Philosophy of Science* 35: 322–54.

LARAUDOGOITIA, J. (1996). 'A Beautiful Supertask', *Mind* 105: 81–3.

LEWIS, D. (1979). 'Counterfactual Dependence and Time's Arrow', *Noûs* 13: 455–76.

MALAMENT, D. (1995). 'Is Newtonian Cosmology Really Inconsistent?', *Philosophy of Science* 62: 489–510.

MAUDLIN, T. (1994). *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics.* Oxford: Basil Blackwell.

MISNER, C., THORNE, K., and WHEELER, J. A. (1973). *Gravitation.* New York: W. H.

Freeman.

NERLICH, G. (1994). *The Shape of Space*. 2nd edn. Cambridge: Cambridge University Press.

NORTON, J. (1999). 'The Cosmological Woes of Newtonian Gravitation Theory', in H. Goenner, J. Renn, J. Ritter, and T. Sauer (eds.), *The Expanding Worlds of General Relativity: Einstein Studies*. Boston: Birkhäuser, vii. 271–322.

—— (2007) 'What Can We Learn about the Ontology of Space and Time from the Theory of Relativity?', in L. SKLAR (ed.), *Handbook of Philosophy of Science*. Oxford: Oxford University Press.

REICHENBACH, H. (1956). *The Direction of Time*. Berkeley: University of California Press.

—— (1958). *The Philosophy of Space and Time* New York: Dover.

ROBB, A. A. (1914). *A Theory of Space and Time*. Cambridge: Cambridge University Press.

RUSSELL, B. (1913). 'On the Notion of Cause', *Proceedings of the Aristotelian Society* 13: 1–26, repr. in *Mysticism and Logic*. Garden City, NY: Doubleday, 1957: 174–201.

SALMON (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

TELLER, P. (1991). 'Substance, Relations, and Arguments About the Nature of Space-Time', *Philosophical Review* 100: 362–97.

—— (1995). *An Interpretive Introduction to Quantum Field Theory*. Princeton: Princeton University Press.

VAN FRAASSEN, B. (1970). *An Introduction to the Philosophy of Time and Space*. New York: Random House.

CHAPTER 35
CAUSATION IN BIOLOGY

SAMIR OKASHA

INTRODUCTION

Biology, as everybody knows, is the study of living organisms. Modern biology is divided into a large number of subdisciplines, each with its own explanatory agenda, research protocol, and specialized vocabulary. These range from molecular biology and biochemistry, which study the molecular and chemical basis of life, to cell biology, embryology, and 'whole organism' biology, which study life at higher levels of organization, to ecology, paleontology, and phylogenetic systematics, which study life over extended spatial and temporal scales. Given this heterogeneity, it might be wondered whether anything useful could be said about causation in biology that would not be so general as to apply to every other science too. Though natural, this sentiment is in fact misplaced. The concept of causation *does* raise distinctive problems in a biological setting, at least some of which find no parallels in the physical sciences. This is particularly true of causal concepts as they feature in evolutionary theory and genetics, which will be the focus of attention here.

Restricting our focus to evolutionary theory and genetics may seem unusual, given the wealth of other biological subdisciplines, but it is a natural choice for at least two reasons. First, Darwinism is arguably *the* grand unifying theory in modern biology, much as Newtonianism was the grand unifying theory in eighteenth- and nineteenth-century physics. (The Russian-American biologist Theodosius Dobzhansky wrote in 1967 that 'nothing in biology makes sense except in the light of evolution', a statement that still holds true today.) So philosophical problems that arise in evolutionary theory (of which genetics is an important part) are in a sense problems for the whole of biology; and causation is the source of many such problems. Secondly, there is an intimate historical link between evolution and genetics on the one hand, and statistics and the methodology of causal inference on the other. Figures such as Francis Galton, Karl Pearson, Ronald Fisher, and Sewall Wright were intimately involved in both enterprises.

The discussion below is organized into four sections. Section 1 is a brief historical survey of the contributions made by biologists to the understanding of causality. Section 2 looks at the role of causal concepts in the theory of evolution. Section 3 discusses Mayr's distinction between proximate and ultimate causation, and the related issue of teleological explanation.

Section 4 looks at causation in genetics, with special reference to the nature–nurture problem.

1. BIOLOGISTS AND THE STUDY OF CAUSALITY: WRIGHT AND FISHER

Aristotle was probably the first person to make significant contributions to both biology and the study of causality; and these two aspects of his work were interrelated. In the early years of the twentieth century, pioneering work on the methodology of causal inference was done by Sewall Wright and R. A. Fisher, both of whom were distinguished evolutionary biologists; like Aristotle, their interest in causality was also closely linked with their biological work. Wright's major contribution was the development of path analysis, a statistical technique for analysing patterns of causal influence between variables, which he put to use in his experiments on breeding and inheritance. Fisher's major contribution was to devise the randomized experiment, widely regarded as the only reliable way of extracting causal information from nature. Though path analysis was temporarily forgotten, both Wright and Fisher have had a lasting influence on the way modern scientists think about causality.

A distinctive aspect of Wright's and Fisher's work was their use of statistics to tackle questions about causality. It is no accident that this innovation was first made by biologists, for the science of statistics (to which Fisher contributed greatly) was originally developed with an eye to biological applications. The concept of correlation, which lies at the heart of descriptive statistics, was first formulated by Francis Galton, a cousin of Darwin, who was interested in the study of heredity; the mathematical definition of correlation was devised by Karl Pearson, pioneer of the newly established discipline of biometry, and one of the founders of modern statistics. Though Wright and Fisher belonged to the rival school of Mendelians, to whom the biometricians were bitterly opposed, they shared the aim of tackling problems of heredity and evolution by using statistical methods, and drew on Pearson's work. However, there was one crucial philosophical difference between Pearson, Wright, and Fisher. Pearson was a staunch logical positivist; he believed that science should dispense with the notion of causation altogether, which he called a metaphysical 'fetish', and replace it with the mathematically well-defined concept of correlation. Wright and Fisher did not accept this view, nor the associated Humean view that causation 'just is' correlation. They regarded causation as conceptually distinct from correlation, and were concerned with how causal inferences could be made from correlational data.

Wright outlined the method of path analysis in a 1921 paper entitled 'Correlation and Causation' and refined it over the 1920s. The method was designed to analyse complex causal systems in which a given effect, or response variable, is believed to be causally influenced by a number of other variables, which may themselves be correlated, or may causally influence each other. Wright used a simple diagrammatic representation of a causal system that he called a 'path diagram'; today, path diagrams are known as 'causal graphs' (see Ch. 14, 'Causal Modelling'). In the path diagram in Fig. 35.1, the effect in which we are interested is denoted 'X'. The three straight arrows leading into X indicate that X is causally influenced by three other variables, A, B, and C; the Greek letters alongside each arrow denote the strength of causal influence along each path, and are called path coefficients. The double-headed

arrows indicate that each of A, B, and C is correlated with each of the other two. (These correlations may be due to the causal influence of other variables, not depicted in the diagram; for example, it is possible that A and B are both causally influenced by a further variable D, and that that is why they correlate. Any path diagram must make a choice about which variables to include and which to omit.)
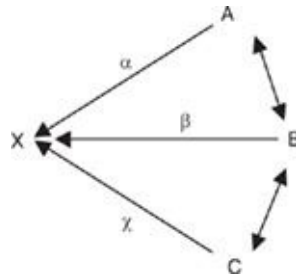
Fig. 35.1 A path diagram.

The aim of Wright's method was to provide a way of estimating the strength of the causal influence running along each path, that is, the path coefficient, from observational or experimental data. The data take the form of measurements of the values of each of the variables in a given population. Thus for example, suppose that X denotes the adult height of a corn plant, A denotes the average height of its two parents, B denotes the number of hours of sunshine received by the plant, and C denotes the amount of fertilizer added to the field where the plant lives. (The path diagram then expresses the hypothesis that a corn plant's height is causally influenced by the height of its parents, the amount of sunshine it receives, the amount of fertilizer it receives, and nothing else.) It is straightforward to measure the values of these variables on a large number of corn plants, and thus to compute the coefficients of correlation between each pair of variables. Path analysis enables us to go a step further, and compute the values of the path coefficients from the observed correlations, thus enabling us to compare the strength of the causal influence transmitted along each path. For example, from a given set of data, the values for the three path coefficients might be $\beta = 0.4$, $\alpha = 0.3$, $\chi = 0.7$, indicating that amount of fertilizer added exerts a stronger causal influence on plant height, compared to the other two variables in the model. (In modern terminology, Wright's path coefficients are simply standardized partial regression coefficients, that is, they measure the statistical association between two variables, controlling for other confounding variables.)

From this brief description of path analysis, one might think that Wright was attempting to do the impossible, that is, to deduce causal relations from observed correlations alone. Indeed, this is how Wright was interpreted by an early critic H. E. Niles (1922). However, Wright (1923) actually took pains to stress that he was *not* trying to do this. On the contrary, the method of path analysis can only be applied once a path diagram is specified, and to specify a path diagram is to advance a set of hypotheses about the causal relations between the variables in question; if this is to be done in a non-arbitrary way, background causal knowledge is essential. Once a path diagram has been specified, path analysis then allows us to deduce the path coefficients from the observed correlations, but this is all on the assumption that the path diagram is correct. As Wright (1921: 559) said, 'the method depends on the combination of knowledge of the degrees of correlation among variables in a system, with such knowledge as

may be possessed of the causal relations. In cases in which the causal relations are uncertain the method can be used to find the logical consequences of any particular hypothesis in regard to them.' This quotation makes it quite clear that Wright was using observed correlations *plus background causal assumptions* to make inferences about the strength of the causal influence between variables of interest, which is very different from trying to deduce causation from correlation alone. In fact, Wright's discussion of path analysis conforms nicely to Nancy Cartwright's (1989) dictum 'no causes in, no causes out', and was methodologically ahead of its time.

It is important to realize that path diagrams and causal graphs depict causal relations between *variables* (or the properties they represent), not between singular events. Thus in Fig. 36.1, X cannot be interpreted as representing the height of one particular corn plant, A the height of its parents, and so on. It is presumably true that the height of any particular plant is causally affected by multiple factors, but this is not the sort of causal relation that path analysis deals with. Rather, path analysis is concerned with causal relations expressed by statements of the form '*differences* in factor A cause *differences* in factor X'; for example, it can help determine whether differences in the amount of fertilizer applied cause differences in the height of plants. Wright fully appreciated this contrast between 'singular' and 'population-level' causation, as it sometimes called today (cf. Hitchcock 1995). He wrote: 'the problem as to the relative importance of heredity and environment in determining the characteristics of a single given individual has no meaning. Both are absolutely essential ... On the other hand, the problem as to their relative importance in determining variation, or *differences*, within a given stock, has a perfectly definite meaning and can often be solved with great ease' (Wright 1921: 250, emphasis in original). This point is crucial for understanding causation in genetics, as we shall see.

Wright's work on path analysis did not have the impact it deserved, despite his fame as an evolutionary biologist. In the 1960s, path analysis was effectively rediscovered by economists and social scientists, leading to the discipline known as 'structural equation modelling' (cf. Blalock 1961; 1964). In recent years there has been considerable interest in causal graphs as tools for investigating causality; see in particular Pearl (2000). Wright's path analysis is the intellectual ancestor of all the recent work on causal graphs, though he would not necessarily have endorsed the direction in which this work has gone.

Fisher developed his ideas on randomized experiments at the Rothamstead Experimental Station in the 1920s and 1930s; his influential books *Statistical Methods for Research Workers* and *The Design of Experiments* appeared in 1925 and 1935 respectively. Fisher argued that causal inferences can safely be drawn from experimental data only if the experiments are appropriately designed; randomization was a key element of correct design. An example can help bring out the logic of the randomized experiment. Suppose a pharmaceutical company claims to have found a drug that helps cure malaria. How might we test this claim? The first step is to find a large population of malaria sufferers willing to participate in a clinical trial. We then divide the population into two groups of equal size; one group receives the drug, the other does not. If the drug really does help cure malaria, we should expect members of the first group to exhibit milder malarial symptoms than members of the control group, after a suitable period of time has elapsed. If on the other hand the drug has no effect on malaria, that is, if the 'null hypothesis' is true, then we should expect no

significant differences between the two groups. Once the result of the experiment is in, we can use a significance test to determine the probability of obtaining that result, if the null hypothesis were true. If this probability is sufficiently low, we are justified in rejecting the null hypothesis, that is, concluding that the drug does indeed help cure malaria.

This brief description of a clinical trial is silent on one key point. How do we divide the population into the two groups in the first place? What determines whether a given individual goes into the treatment group or the control group? Fisher's key idea was that this must be done at *random.* For example, the experimenter could put the names of all the individuals into a hat, mix them thoroughly, and draw them out one by one, allocating alternate names to each group. Alternatively, a randomizing device such as a roulette wheel or a coin-flip could be used; or more simply, a published table of random numbers. Only if the division into treatment and control groups is done randomly can causal inferences be legitimately extracted from the experiment, Fisher argued.

Why did Fisher place such a premium on randomization? His reasons were essentially two (cf. Fisher 1935; Shipley 2000). First, he argued that significance testing is only theoretically justified if the experiment is randomized. Recall that the significance test involves calculating the probability that the actual outcome would have occurred, if the null hypothesis were true (in our example, if the drug has no curative effect). This probability is calculated by assuming that, if the null hypothesis is true, then the result of the experiment 'will be wholly governed by the laws of chance' (Fisher 1935: 20). This assumption is only valid, Fisher argued, if the division of the population into treatment and control has been made at random. For otherwise, we cannot infer that a member of the treatment group and the control group have equal probability of recovering from malaria, given the truth of the null. In Fisher's words: 'the simple precaution of randomization will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged' (ibid. 21).

Secondly and more controversially, Fisher argued that randomization helps ensure that any observed correlation between experimental treatment and outcome that passes the test of significance will be *due* to the treatment. It is obvious that correlation is not a general guarantee of causation. But in the context of a randomized experiment, Fisher argued that the inference from correlation to causation *can* be safely made. In our example, if patients in the treatment group recover better than those in the control group, and the significance test has ruled out chance as the explanation, then we can infer that the treatment caused the recovery. Obviously this inference is not *deductively* valid, but it is a reasonable one, Fisher argued. For while it is possible that the difference in recovery between the treatment and control groups was due to some factor other than the drug, for example, to the preponderance of a particular genotype in one of the groups, the fact that the population was divided into the two groups at random makes this unlikely. Given randomization, there is no reason to think that any particular confounding factor will be over-represented in either group. So a significant difference between the groups is presumptive indication of the causal efficacy of the treatment, according to Fisher.

Fisher's work on experimental design has been extremely influential, across the natural and social sciences. His thesis that the randomized experiment is the only reliable way to make causal inferences, and that purely observational data is of no use for this purpose, has achieved orthodoxy in many disciplines, for example, epidemiology. (Famously, Fisher used this thesis

to dispute the link between smoking and lung cancer, arguing that the link had not been established because randomized experiments had not been performed.) But despite Fisher's influence, there is an ongoing debate, in both statistics and philosophy of science, over the exact significance of randomization. Not all theorists agree with Fisher than the randomized experiment is the uniquely privileged way of acquiring causal knowledge. For example, Shipley (2000) notes that randomized experiments are not feasible in many areas of science, but that causal inferences are nonetheless made. Howson and Urbach (1989) point out that even with randomization it is overwhelmingly likely that the treatment and control groups will differ in *some* respect (given the large number of possible 'respects'), which for all we know might influence the experimental outcome. So it is quite wrong, they claim, to regard randomization as a universal panacea for the problem of confounding factors, as many scientists do. So despite Fisher's legacy, deep philosophical issues over the justification for randomization remain unresolved.

2. Causation and the Theory of Evolution

Darwin's theory of evolution by natural selection was an attempt to explain two phenomena: adaptation and diversity. 'Adaptation' refers to the fact that most organisms possess features that seem incredibly well suited for survival in their particular environment. 'Diversity' refers to the fact that living organisms are not all alike, but are clustered into relatively discrete types, or kinds. Traditionally these phenomena were explained by invoking an intelligent designer, or deity, but Darwin sought a naturalistic alternative. He argued that the process of natural selection, or differential survival and reproduction of variant organism-types, provided the answer. Natural selection will operate on any population of organisms that satisfies three conditions. First, the organisms must vary with respect to some of their traits, or characters; secondly, some variants must leave more offspring than others; and thirdly, offspring must tend to resemble their parents with respect to the traits in question. Over time the population's composition will then change, as the fitter variants supplant the less fit. Darwin argued that given enough time, this process will lead organisms to evolve adaptations that are closely tailored to the demands of their environment; since environments differ, this in turn will produce diversity.

Darwin's theory lies at the heart of modern biology, helping to make sense of countless phenomena that would otherwise seem puzzling. It is routine for biologists to seek Darwinian explanations (also known as 'adaptive explanations') for organismic features they are interested in. Thus suppose, for example, that a lepidopterist observes an unusual wing-coloration pattern in a given species of butterfly. She will probably assume that the pattern is there for a reason, that is, that it bestows some advantage on the butterflies that have it. For example, the pattern might help make a butterfly less visible to predators, or it might help screen out harmful UV radiation. Discovering the adaptive function of a trait is not easy, and it is of course possible that it has none, that is, that it did not evolve by natural selection. But for complex traits, for example the vertebrate eye, which seem clearly beneficial for the creatures possessing them, it is generally assumed that there is a Darwinian explanation for how the

traits got there. Spelled out fully, the explanation would say that in an ancestral population, organisms possessing the trait (or some rudimentary precursor of the trait) had a survival advantage for some reason over those not possessing it, and that the trait was heritable, that is, passed down from parents to offspring.

Many scientific explanations are causal, and Darwinian explanations are no exception. To explain a trait's prevalence in a population by natural selection is to say what *caused* the trait to evolve and be maintained in the population. So natural selection is a potential cause of evolutionary change, and Darwinian explanations are causal hypotheses about the evolution of particular traits. However, the logic of Darwinian explanation is somewhat different from that of other causal explanations in science. An important distinction due to Elliott Sober (which he credits to Richard Lewontin) helps show why. Sober (1984) distinguishes between what he calls 'developmental' and 'selectional' explanations. To illustrate the difference, suppose we are trying to explain why a class of schoolchildren is academically outstanding. One possible explanation would look at each child in the class individually, citing the reasons for their exceptional performance—which might include their dedication to study, their stimulating home environment, etc. This is a developmental explanation—it cites the causal factors that have led each child to develop into a high achiever; by aggregation, this yields an explanation of why the class as a whole is outstanding. But a second, quite different explanation is that in order to get into the class in the first place, children need to have scored above 95 per cent in the previous year's examination; as a result, the class contains only outstanding children. This is a selectional explanation. It explains why the class is academically outstanding by describing the selective process used to determine class membership, rather than by explaining the abilities of the particular children in the class.

Importantly, the developmental and selectional explanations are complementary rather than competing. This is because, as Sober points out, the two explanations have slightly different *explananda*. The former explains, of each child in the class, why the child developed into a high achiever, rather than that *self-same child* developing into a low achiever. The latter explains, of the class as a whole, why it contains high achieving children, rather than containing *other* children who are low achievers. Also importantly, both explanations are causal. The developmental explanation tells us what caused each child in the class to become a high achiever. The selectional explanation tells us what caused the class as a whole to be *composed* of high achievers. There is a perhaps a sense in which the developmental explanation is more causally fundamental—for it is an explanation 'from the bottom up'. But this should not lead us to deny that the selectional explanation is causal too. After all, the class could have been composed of low achievers, or of a mixture of high and low achievers, so *something* must have caused it to have the composition it actually has; that thing is the selective process by which entry to the class was determined. Had a different selective process been employed, the composition of the class would have been different. So although the selectional explanation does not tell us what caused any individual child to turn into a high achiever, it is causal nonetheless.

In many areas of biology, we find developmental-type explanations. For example, a geneticist might explain why someone suffers from Down's syndrome by citing the fact that they have three copies of chromosome 21. The logic of this explanation is easy to understand —it simply identifies the (main) causal factor that led this particular individual to develop

Down's syndrome. But Darwinian explanations are not like this; they are selectional. A standard Darwinian explanation does not aim to explain a fact about a particular individual but rather a fact about the composition of a *population* of individuals. For example, consider a Darwinian explanation of why giraffes are so tall. The explanation would say that in an ancestral giraffe population, there was variation in height; taller giraffes had a survival advantage over less tall ones (e.g. because they could reach higher foliage), and height was a heritable trait; so over time, the population grew taller and taller. As Sober stresses, this explanation tells us what caused the composition of the population to change over time; but it does not tell us what caused any individual giraffe to grow tall rather than small. (To explain this, we would need to cite the genetic and environmental factors affecting that particular giraffe; this would be a developmental explanation.) Again, it is important to see that Darwinian explanations *are* causal, even though their logic is different from that of other causal explanations in science.

The causal structure of Darwinian explanations can be further elucidated by examining a second distinction, also due to Sober (1984), between 'selection of' and 'selection for'. To see this distinction, recall the second of Darwin's three conditions for evolution by natural selection. The condition says that some variants in the population must leave more offspring than others, that is, reproduction must be differential. The rationale for this condition is obvious—if all organisms leave the same number of offspring, natural selection cannot operate. But the condition as stated contains an ambiguity. Is the idea that an organism's reproductive output (fitness) must *causally* depend on which traits it has, or can the dependence simply be statistical? In Sober's terms, when a given trait causally affects organismic fitness there is selection *for* that trait; but if the trait merely correlates with fitness, without causally affecting it, there is selection *of* the trait. Importantly, selection of and selection for can both lead to evolution by natural selection; what matters for evolution is *that* some variants leave more offspring than others, not why they do so. (This is why the ambiguity in the 'differential reproduction' condition usually does no harm.) But if we wish to understand the ecological basis of natural selection, or to understand the adaptive function of evolved traits, Sober's distinction is crucial.

An example can help illustrate the selection of/for distinction. Suppose that body size and brain size are positively correlated in a given primate species, for reasons to do with embryological development. Body size causally affects fitness—larger organisms are fitter than smaller ones—but brain size is selectively neutral. Therefore, there is selection *for* being large-bodied but not *for* being large-brained. However, since body and brain size are correlated, there will be selection *of* large-brained organisms. Organisms with large brains will leave more offspring than those with smaller brains, not *because* they are large-brained, but rather because they are large-bodied and large bodies happen to go along with large brains. (Biologists sometimes express this by saying that body size is directly selected, brain size indirectly selected; this is an alternative way of capturing Sober's distinction.) Presuming that body and brain size are both heritable, both will evolve by natural selection—the average value of both traits in the population will increase over generations. So if we simply wish to predict the outcome of natural selection, the selection of/for distinction makes no difference. But to understand the causal basis of the observed evolutionary change we need to know which traits have been selected for, and why.

The concept of selection for a trait helps bring out a further important point about causality as it relates to evolutionary theory. Recall the distinction between 'population-level' and 'singular' causation from the previous section; this is the distinction between such statements as 'smoking causes lung cancer' and 'Mr Jones' heavy smoking caused him to get lung cancer.' So far as Darwinian explanations are concerned, the relevant sort of causality is population-level, rather than singular. To see this, consider what exactly it means to say that body size was 'selected for', in the previous example. It means that *differences* in organisms' body size caused *differences* in their reproductive output—which is a statement of population-level causality. (Note that 'smoking causes lung cancer' could easily be paraphrased as 'differences in quantity smoked cause differences in susceptibility to lung cancer'.) Singular causal relations are not relevant to Darwinian explanations, given that such explanations aim to explain facts about populations, not individuals. The complete causal explanation of why any individual produced the number of offspring it did will obviously be extremely complex; fortunately, evolutionary biologists do not need to concern themselves with it. What matters, vis-à-vis Darwinian explanation, is to find those traits that causally affect fitness, where this is understood to mean that differences in those traits cause fitness differences in a given population. Individual organisms get no mention in explanations of this sort, so the myriad of causal factors affecting any individual's reproduction, which would probably be impossible to determine anyway, is beside the point.

3. PROXIMATE VERSUS ULTIMATE CAUSATION, TELEOLOGY, AND NATURAL SELECTION

In a famous paper entitled 'Cause and Effect in Biology' (1961), Ernst Mayr drew a distinction between what he called 'functional' and 'evolutionary' biology. Functional biologists are concerned with how organisms and their parts work; the primary question they ask is 'how'. Thus, for example, a physiologist might ask how an organism manages to regulate its body temperature; an immunologist might ask how the immune system manages to attack antigens without hurting the host; a cell biologist might ask how the timing of cell division is controlled. Evolutionary biologists, by contrast, are interested in a fundamentally different sort of question. They want to know *why* organisms exhibit the features they do; the primary question they ask is 'why', not 'how'. When an evolutionary biologist studies an organism, the aim is to understand the adaptive significance of the organism's traits, that is, to understand why the process of natural selection has led to the evolution of those traits rather than others. What evolutionary advantage did the traits bestow, that led natural selection to favour them?

Mayr argued that both types of biology deal with causal questions, but of different types. Functional biology is concerned with 'proximate' causation, evolutionary biology with 'ultimate' causation. The functional biologist aims to identify causal mechanisms at work in modern organisms that produce the biological features they are interested in. Mayr (1961: 1502) wrote: 'the chief technique of the functional biologist is the experiment, and his approach is essentially the same as that of the physicist and the chemist'. Evolutionary biologists aim to identify the causal-historical pathway by which modern organisms came to

possess the features they do, and the reasons for the evolution of those features; experimentation is generally of little use in this quest.

The dichotomy between the two types of biology is not absolute. To understand how modern organisms evolved, one needs to know at least something about how they actually work; so a certain amount of functional biology is a prerequisite for any evolutionary inquiry. Conversely, an evolutionary perspective is often useful for studying modern organisms, primarily because it licenses the assumption that most features of organisms are likely to somehow benefit them. So there is scope for interplay between the two sorts of biological inquiry. Despite this, Mayr is quite right that evolutionary biologists are concerned with a different *sort* of question than are biologists of other stripes. (However, Mayr's use of the expression 'functional biology' to denote non-evolutionary areas of biology is non-standard and potentially quite misleading. For Darwinian explanations are sometimes referred to as 'functional explanations'; and traits that evolved by natural selection are often said to have a 'function' (or an 'adaptive function'). So Mayr's terminology is not ideal, but his distinction is an important one.)

Mayr's characterization of evolutionary biology as concerned with 'ultimate' causation leads naturally to the question of teleology, or goal-directedness. Teleological and purposive language is widely used in biology, both past and present. We say that the squirrels hoard food *in order* to survive the harsh winter months; that plants *try* to get as much sunlight as possible; that the *function* of the kidneys is to cleanse the body of waste products; and that a finch's beak is *for* cracking seeds. This way of speaking is extremely natural. Many organisms behave as if they are consciously trying to achieve some goal (though the majority are presumably not capable of conscious thought at all); and many bodily organs look as if they have been deliberately designed by a conscious agent to perform some task (though we know that they have not). The use of teleological language in biology stems from a pre-Darwinian worldview, according to which living organisms were created by God. Given this worldview, there is a clear rationale for describing organisms in teleological terms. We routinely describe the products of human design teleologically, as when we say that the fan is *supposed* to prevent the engine overheating; this means that a human designer constructed the fan with that purpose in mind. If living organisms are the products of God's design, it is natural to describe them in teleological terms too.

On the modern Darwinian worldview, the use of teleological language in biology becomes more problematic. For Darwinism teaches us that living organisms are the product of evolution by natural selection, so were *not* created by God. However the process of natural selection, when it operates cumulatively over many generations, produces organisms that *look* as if they have been designed by a conscious intelligence, so perfectly do their adaptations fit the demands of their environment. (Think for example of a stick insect, which matches the surrounding foliage unbelievably well.) Therefore, although Darwinism obviously removes the traditional rationale for describing organisms and their bodily organs in teleological terms, there is a sense in which it supplies a different one. As noted above, organismic traits and behaviours that have been shaped by natural selection are often said to have adaptive functions; a trait's adaptive function is the thing that it does in virtue of which it was selected for in the past. 'Function' is of course a paradigmatically teleological notion, as is the related notion of an organismic trait being *for* something.

In a recent article, Colin Allen (2003) writes: 'opinions divide over whether Darwin's theory of evolution provides a means of eliminating teleology from biology, or whether it provides a naturalistic account of the role of teleological notions in science' (p. 1). It is easy to see where this difference of opinion stems from. Those who hold that Darwinism eliminates teleology stress that natural selection produces the *appearance* of design in nature, rather than the real thing; one of Darwin's main achievements, they argue, was to show that the teleology in nature is only apparent. Those who hold that Darwinism naturalizes, rather than eliminates, teleology stress that the notion of a trait's adaptive function is a key part of Darwinian explanations, so Darwinism clearly doesn't eliminate all teleological language. Moreover, for many traits, the functional attributions licensed by Darwinism coincide with those that pre-Darwinian biologists would have made. For example, pre-Darwinian biologists claimed that the function of the heart was to pump blood. Darwinians agree with this statement, so long as 'function' is understood as adaptive function—for it is a reasonable assumption that ancestral hearts were selected for their ability to pump blood. The difference between these two viewpoints is largely terminological. Mayr (1961) recommended the word 'teleonomy' for the scientifically respectable sort of teleology embodied in the Darwinian notion of function, while reserving 'teleology' for the unrespectable sort that Darwinism eliminates, but his recommendation did not catch on.

Causality is relevant to the debate over teleology for this reason: traditionally in philosophy, teleological explanation has been contrasted with causal-mechanistic explanation. The idea behind this contrast is that in causal-mechanistic explanation one explains a given feature by reference its causes; while in teleological explanation, one explains a given feature by reference to its effects—as, for example, when one explains the operation of a thermostat by saying that it ensures the room is kept at a constant temperature. (The relation between causal and teleological explanation is a contentious issue; some philosophers have argued that any legitimate teleological explanation, fully spelled out, must in fact *be* causal. Whether this is correct does not matter here; the point to note is just that teleological and causal explanation have traditionally been opposed.) Those biologists who argue that Darwinism has eliminated teleology from biology emphasize, correctly, that Darwinian explanations are causal, and that natural selection is 'blind' to what will happen in future. Since teleological explanations, on the standard characterization, explain current phenomena by reference to their future effects, these biologists regard teleology as the antithesis of Darwinism. They argue that Darwin's achievement was precisely to provide a causal-mechanistic explanation of phenomena that had previously been thought to require teleological explanation.

Disputes about how the word 'teleology' should be used aside, the point that natural selection is blind to the future is an important one. To appreciate the point, recall how the Darwinian process works: the fittest variants in a population enjoy greater reproductive success than their less fit counterparts, and transmit their fitness-enhancing traits to their offspring. Crucially, variants are only selected if they provide a *current* fitness advantage in the *current* environment. There is no way that selection can favour traits that are currently of no benefit, but that might become beneficial in the future; natural selection has no foresight. (It is important to remember that 'natural selection' is just a metaphor for the process of differential reproduction; no one literally does any selecting.) Also important is the fact that the process of genetic mutation, which creates the variation on which natural selection acts, is

random, or undirected. Favourable mutations do not arise *because* they are favourable, but rather by chance. For example, imagine a gene A that mutates to B with a certain probability. If the environment of the species changes so that the B gene becomes greatly beneficial, this will *not* increase the rate of mutation from A to B. Note that the randomness of mutation is an empirical truth (and counterexamples are occasionally alleged); by contrast, natural selection's lack of foresight is a conceptual truth that follows from the logic of Darwinism. Taken together, the two points serve to emphasize that Darwinian selection is a bona fide causal process, and in no sense forward-looking.

4. CAUSATION IN GENETICS

Genetics is a highly important area of modern biology. Modern molecular genetics began in 1953 with Watson and Crick's discovery of the structure of DNA, the material from which genes are made. In the years that followed, the mechanisms by which genes are copied, via DNA replication, and by which they specify the amino acid sequence of proteins, via transcription and translation, were rapidly uncovered. More recently, developmental molecular genetics has made great strides forward in explaining how a single fertilized zygote can give rise to a complex organism with differentiated cell and tissue-types, though much remains to be found out.

Importantly, genetics began life in the pre-molecular age. Classical genetics was a flourishing field in the 1920s and 1930s, long before anything about DNA was known. Unlike in modern genetics, the gene of classical genetics was a purely theoretical entity, posited in order to explain observed patterns of inheritance, with no assumptions made about its physical reality. Classical geneticists were primarily interested in studying the transmission of genes, and phenotypic traits, across generations; their methodology, like Mendel's, was to analyse the results of hybridization experiments. *How* genes produced phenotypes was not something they were concerned with; it was simply taken for granted. The precise relation between classical and molecular genetics is a subtle issue; suffice to say that it is unclear whether 'gene' means exactly the same in both disciplines.

Causation is central to a number of important controversies within genetics. One concerns the relation between genes and phenotypic traits. It is common to hear it reported that scientists have discovered a 'gene for' a certain trait, for example, bipolar depression. But what exactly does this mean? Is it really plausible that a complex phenotype, especially a behavioural one, can be attributed to the action of genes alone? Critics see here the threat of genetic determinism, the idea that genes rigidly determine all aspects of phenotype. Particularly when the species in question is our own, this is an unwelcome prospect. However, the 'gene for' locution does not in fact carry the implication of genetic determinism, despite what is often thought, for there is an implicit relativity to background conditions. The vast majority of traits are causally affected by many genes and many environmental factors. Even traits regarded as paradigmatically 'genetic' are typically dependent on the environment to some degree. (For example, phenylketonuria, a disorder of the metabolism in humans, was long regarded as purely genetic; it is now known that development of the condition also

depends on diet.) To say that there is a 'gene for' a trait does not mean that the gene produces the trait all on its own; rather, it means that given 'normal' background conditions—including genetic background—the gene in question makes a difference to whether the trait appears or not. This does not imply that development of the trait is inevitable, given the gene, or that the environment has no role to play. Correctly understood, the 'gene for' locution is thus much less problematic than critics have charged.

It is also important to realize that virtually all Darwinian explanations are implicitly committed to the existence of a 'gene for' the trait whose evolution they are concerned with. We have seen that heritability, or parent–offspring resemblance, is crucial to evolution by natural selection. Even if a trait is strongly selected for, this will not produce a cross-generational evolutionary response unless the trait is heritable. The reason that most phenotypic traits are heritable is that parents and offspring share genes—parents transmit their DNA to their children. (Though traits can be heritable for non-genetic reasons too.) So when a biologist gives a Darwinian explanation of a trait's prevalence in a population, or its evolution over time, there is an implicit assumption that the trait is causally influenced by genes, that is, that trait differences are partly caused by genetic differences, even though the genes in question can rarely be identified at the molecular level. This explains why evolutionary biologists feel justified in speaking of a 'gene for' a trait even when no actual molecular gene has been discovered. Their justification is that if the trait were not causally affected by one or more genes, it would not be heritable, in which case it could not evolve by natural selection; thus the 'gene for' assumption is only as controversial as the assumption that the trait is a Darwinian adaptation. Of course, we may be unsure whether a given trait *is* an adaptation; the point is that *if* it is then it must have a genetic basis, at least in part.

The subject called 'behaviour genetics' has long been the seat of methodological controversy. Behaviour geneticists study human behavioural traits, especially cognitive traits such as IQ. Their aim is to 'disentangle nature and nurture', that is, to determine whether differences between humans are due to genetic factors, environmental factors, or both. Their main methodological tool is *heritability analysis,* a statistical technique for determining how much of the variability of a trait, in a given population, is attributable to genetic differences. (A trait's heritability is thus a number between 0 and 1, i.e. the proportion of the total variance that is due to genetic variance.) Behaviour geneticists routinely produce statements of the form 'the heritability of height is 0.8', 'the heritability of IQ is 0.6', and so on; the implication is often given that the higher a trait's heritability, the more 'genetic' it is. The controversy surrounding behaviour genetics is multifaceted. Opponents have argued that traits such as IQ are very poorly defined, that certain behaviour geneticists have suspect political motivations, and that heritability analysis is methodologically flawed. In particular, the inference from 'high heritability' to 'genetic causes' has often been contested, as has the intellectual value of producing heritability estimates in the first place.

It is certainly true that heritability analysis can tell us nothing about the causes of a trait in a particular individual. (Behaviour geneticists have rarely claimed otherwise, but confusion on this point is rife in popular discussions.) For heritability is an irreducibly population-level parameter. A given trait may have a high heritability in one population and a low heritability in another, or in the same population at a different time. If a population's environment is made more homogenous, then the heritability of every trait will automatically increase—for a

greater fraction of trait differences will be attributable to genetic differences; in the limit, if the environment were made perfectly homogenous, every trait would have a heritability of one. Conversely, if a population is made more genetically uniform, then the heritability of every trait declines; in the limit, in a purely clonal population every trait has heritability of zero. This drives home the point that heritability estimates are population-relative; they tell us nothing about individuals. A given individual could be moved from one population to another, in which the heritability of height (for example) was quite different; obviously, the causal factors affecting that individual's height would be unchanged.

The distinction between singular and population-level causality is crucial here. Heritability analysis, in so far as it licenses any causal inferences, pertains only to population-level causality. When a behaviour geneticist asserts that height is 60 per cent genetic and 40 per cent environmental, this does *not* mean that in any particular individual, genes are responsible for 60 per cent of their height and the environment for the remainder; such an assertion would obviously be meaningless. Rather it means that in a given population, genetic differences are responsible for 60 per cent of the variability in height, environmental differences for the remaining 40 per cent; this says nothing about the cause of any individual's height. Recall the quotation from Sewall Wright in sect. 2 above, concerning the impossibility of apportioning causal responsibility between heredity and environment for any individual, but the 'ease' of doing so for differences in a population.

However, some critics have argued that heritability analysis cannot even license statements about population-level causality, for the whole idea of partitioning a trait's variability into genetic and environmental components is suspect (cf. Sarkar 1998). Their objection is that gene–environment interaction is probably ubiquitous, and undermines the significance of heritability estimates. Gene–environment interaction means that genetic and environmental factors do not combine additively to determine trait-value; a given genetic factor may affect the trait positively in one environment, but negatively in another. For example, a given gene might make children taller if they are raised on a high-protein diet, but smaller if they are raised on a low-protein diet. (Note that 'gene–environment interaction' involves the technical statistical notion of interaction; it is *not* just the platitude that genes and environment both contribute to the development of the phenotype.) The extent of gene–environment interaction is an empirical issue, on which biologists divide. Opponents of heritability analysis tend to think that such interaction is the norm, and that it renders meaningless the idea of attributing a determinate fraction of a trait's variability to genetic causes. This tricky issue is unlikely to be resolved soon.

Finally, it is worth looking briefly at a recent discussion in philosophy of biology involving both genetics and causality. Proponents of 'developmental systems theory' (DST) argue that much current biology is too gene-centric: DNA is regarded as a 'master molecule' that contains the genetic information specifying how the organism will develop (cf. Oyama, Griffiths, and Gray 2001). This idea is obviously not entirely wrong, but proponents of DST argue that it is both skewed and excessively simplistic. They object to the idea that genes contain 'information', arguing that DNA is simply one of the many causal factors responsible for organismic development. They point out that organisms inherit more from their parents than nuclear DNA alone; cytoplasmic elements, organelles, and ecological resources (e.g. birds' nests) are all inherited too, and are crucial for normal development. Therefore, genes

are not uniquely responsible for the reliable transmission of organismic form across generations; they should be treated as just one developmental resource among many, not a repository of information about how to build an organism.

The issues raised by DST, in particular the critique of genetic information, are difficult ones, with deep historical roots. (Developmental biologists have traditionally been wary of the notion of genetic information, and of the hegemony of genetics in biology.) One philosophically interesting aspect of the DST approach is its plea for 'causal democracy', that is, for treating all the causal factors responsible for organismic development equally, rather than privileging genes (cf. Oyama 2000). The idea of causal democracy seems reasonable; if a given phenomenon is known to have multiple causes, it makes sense to accord all of them equal status, at least until the greater importance of one has been empirically demonstrated. However, at times the advocates of DST have defended the notion of causal democracy in a suspiciously a priori way, as if according parity to the various factors responsible for organismic development is an absolute methodological imperative, rather than something responsible to empirical data. I suggest that this is a mistake. Even if the notion of genetic information cannot be made philosophically respectable, it is still possible that genes turn out to play far *more* important a role in explaining organismic development than other causal factors. Whether this is so is something that we will only know when molecular developmental genetics has run its course.

FURTHER READING

Bill Shipley's book *Cause and Correlation in Biology* (2000) contains a good discussion of Wright's work on path analysis and Fisher's work on randomized experiments, including the relation of their work to modern causal graph theory. Elliott Sober's *The Nature of Selection* (1984) is the best philosophical discussion of causal concepts in evolutionary theory; this is where Sober introduces the selection of/for distinction and the distinction between variational and selectional explanation. Two good biological works discussing causation in relation to evolution are Michael Wade and Susan Kalisz's paper 'The Causes of Natural Selection' (1990) and John Endler's book *Natural Selection in the Wild* (1986). A useful collection of papers on teleological concepts in biology is Colin Allen, Marc Bekoff, and George Lauder (eds.), *Nature's Purposes* (1998); also useful is David Buller (ed.), *Function, Selection and Design* (SUNY, 1999). On causation and genetics, Sahotra Sarkar's book *Genetics and Reductionism* (1998) provides a detailed discussion of most of the important issues, including the limitations of heritability analysis. Elliott Sober's paper 'Apportioning Causal Responsibility' (*Journal of Philosophy* 1988: 303–18) examines the problem of quantifying the relative importance of different causal factors, with particular reference to genetics. The papers in Susan Oyama, Paul Griffiths, and Russell Gray's edited collection *Cycles of Contingency* (2001) defend the developmental systems theory perspective; a good critique is provided by Peter Godfrey-Smith in 'Explanatory Symmetries, Preformation and Developmental Systems Theory' (2000).

REFERENCES

ALLEN, C. (2003). 'Teleological Notions in Biology', *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/teleology-biology/*>*, accessed 23 March 2009.

—— BEKOFF, M. and LAUDER, G. (eds.) (1998). *Nature's Purpose*. Cambridge, Mass.: MIT.

BLALOCK, H. (1961). 'Correlation and Causality: The Multivariate Case', *Social Forces* 39: 246–51.

—— (1964). *Causal Inferences in Nonexperimental Research*. Chapel Hill, NC: University of North Carolina Press.

BULLER, D. (ed.) (1999). *Function, Selection and Design.* New York: SUNY.

CARTWRIGHT, N. (1989). *Nature's Capacities and Their Measurement*. Oxford: Oxford University Press.

ENDLER, J. (1986). Natural Selection in the Wild. Princeton: Princeton University Press.

FISHER, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

—— (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.

GODFREY-SMITH, P. 'Explanatory Symmetries, Preformation and Developmental Systems Theory', Philosophy of Science (Proceedings) 67: S322–31.

HITCHCOCK, C. (1995). 'The Mishap at Reichenbach Fall: Singular versus General Causation', *Philosophical Studies* 78: 257–91.

HOWSON, C., and URBACH, P. (1989). *Scientific Reasoning: The Bayesian Approach*. La Salle, Ill.: Open Court.

MAYR, E. (1961). 'Cause and Effect in Biology', *Science* 134: 1501–6.

NILES, H. E. (1922). 'Correlation, Causation and Wright's Theory of "Path Coefficients"', *Genetics* 7: 258–73.

OYAMA, S. (2000). 'Causal Democracy and Causal Contributions in Developmental Systems Theory', *Philosophy of Science (Proceedings)* 67: S332–47.

—— GRIFFITHS, P., and GRAY, R. (eds.) (2001). *Cycles of Contingency*. Cambridge, Mass.: MIT.

PEARL, J. (2000). *Causality*. Cambridge: Cambridge University Press.

SARKAR, S. (1998). *Genetics and Reductionism.* Cambridge: Cambridge University Press.

SHIPLEY, B. (2000). *Cause and Correlation in Biology*. Cambridge: Cambridge University Press.

SOBER, E. (1984). *The Nature of Selection.* Chicago, Ill.: University of Chicago Press.

WADE, M., and KALISZ, S. (1990). 'The Causes of Natural Selection', *Evolution* 44: 1947–55.

WRIGHT, S. (1921). 'Correlation and Causation', *Journal of Agricultural Research* 20: 557–85.

—— (1923). 'The Theory of Path Coefficients: A Reply to Niles's Criticism', *Genetics* 8: 239–55.

CHAPTER 36
CAUSATION IN THE SOCIAL SCIENCES

HAROLD KINCAID

BECAUSE of the obstacles to experimentation and because of the complexity of the social world, the social sciences present fertile grounds for investigating issues surrounding causation. The issues surrounding experimentation and social complexity are taken up in Chapter 14 (Causal Modelling), Chapter 22 (Causation and Observation), and Chapter 30 (Causation and Reduction), and I will not discuss those topics in any detail. However, much else is left to discuss, more than can be fully covered in one chapter. My goal then is to sketch a number of issues and only secondarily to argue for particular positions.

I approach the issues discussed below with some general background assumptions that frame the issues and are also supported I think by the topics discussed. Those assumptions concern the nature of causal claims in general, more specifically, questions about the extent to which our understanding of causation can be perfectly general. In the past the philosophical project about causation was often the search for necessary and sufficient conditions for the use of the term 'cause' tested against intuitions—that would be a perfectly general account. However, if that project is rejected on the grounds that the metaphysics and epistemology of causation have to be informed by and are continuous with our scientific knowledge of the world, then we should expect things to be more complicated. In particular, we should expect diverse domains to run into rather different issues as the notion of cause gets tied down to the concrete empirical issues of the science at hand. While there may no doubt be many useful general things to say about causation, the full details about causation are likely to depend importantly on specific facts about the way the world works and how evidence is obtained in any given domain.

In what follows I present a number of issues about the ontology and epistemology of causation in the social sciences. The general theme will be that these issues cannot be decided in the abstract but must pay careful attention to the empirical presuppositions made and the kinds of evidence for them.

The topics considered fall into three classes:

(a) Questions about the scope of causal claims in the social science: it has been widely argued that causal claims in the social sciences must be qualified *ceteris paribus* and equally widely doubted that such claims are meaningful and testable. Others have argued that causal notions do not apply to social phenomena. For example, many have argued that aggregate or macrosociological causal claims are incoherent or at least in need of individualist mechanisms. These are the topics of sect. 1.

(b) Questions about specific types of causes invoked by social scientists. Social scientists

invoke various kinds of causes that are in need of clarification and justification. Teleological causes are one such case with a long history of controversy. Social scientists also distinguish between necessary, sufficient, conjunctional, structural, etc. causes. In all these cases it is unclear how these notions are to be understood, applied, and related to core notion of cause. Section 2 discusses these issues.

(c) Questions concerning how causation works in the social world and what that entails for finding evidence about causation. Widely used causal modelling techniques in the social sciences arguably make implicit assumptions about the nature of social causation that fit poorly with some or much social science research. This is interesting in its own right, but also raises important issues about the validity of case study and small N comparison research.

1. THE SCOPE OF CAUSAL CLAIMS IN THE SOCIAL SCIENCES

Causal explanations in the social sciences face an obvious problem. The social sciences seldom have a full accounting of relevant causes and interfering factors. It has thus seemed to many that such claims as 'Increased interest rates cause decreased investment' must be implicitly qualified with a *ceteris paribus* clause. A large literature has thus emerged trying to clarify such claims and analysing what they entail for the scientific status of the social sciences.

Many have tried to defend the legitimacy of such claims by spelling out their truth conditions (Hausman 1992; Pietroski and Rey 1995). There are both specific and general problems with these approaches. For each proposal, it has been easy enough to find serious counterexamples (see Earman and Roberts 1999). Moreover, there are various reasons to think the entire enterprise misguided from the start. First, the appeal to *ceteris paribus* clauses is motivated by the apparent inability to list all the possible interfering factors. That motivation seems in essential tension with the goal of identifying truth conditions—for if we have the latter it looks like *ceteris paribus* clauses would be eliminable in favour of a strict regularity (Earman, Roberts, and Smith 2002). Second, there is no good reason to think that we must first have a truth conditional semantics for claims to be meaningful. We might have some other kind of semantics; not every scientifically respectable claim has explicit truth conditions.

A less direct but popular approach (building on Cartwright 1983) has been to argue that our best science—physics—makes *ceteris paribus* claims and thus that there is no inherent problem with the social sciences doing so as well. A large critical literature has emerged around various issues lurking here. For example, Earman and Roberts and Earman et al. deny that it is *ceteris paribus* 'all the way down' in physics and argue that this is just as well, because *ceteris paribus* generalizations are untestable and provide no coherent notion of laws.

Despite the volume of literature, it seems that participants to the debate talk past one another, especially concerning the core issue here—what is the epistemic status of causal claims in the social sciences or special sciences more generally? The problem arguably comes from merging the issue of well-confirmed causes with separate issues about laws and explanations (see Kincaid 1996: 91–7; Woodward 2002; Kincaid 2004).

Take, for instance, Earman and Roberts's claim that there are no methods for testing *ceteris*

paribus claims, despite various proposals. A natural method for confirming causal claims in the face of possible confounders is to provide repeated independent evidence that confounders can be handled—their presence can be identified in ways that leave us good reason to think the alleged causes are in fact operative. Earman et al. argue that such methods are too weak, because they will count as laws what clearly are not (e.g. 'all white substances are safe to eat' because in each case where they are not we can explain what confounds). However, this has no sting unless the goal is to confirm a law rather than to confirm a cause. We can indeed help confirm the claim that a given substance does not cause ill health in normal human beings by showing that when ill health results, the causal processes are other than normal.

While causal claims arguably entail that laws of some sort exist, it is not the case that we have to have the laws on the books to confirm them or a semantics for *ceteris paribus* laws, if there are any such things (Glymour 2002). Causal claims in the social sciences are *ceteris paribus* claims only in that we know that we can be more or less successful in ruling out confounding and interfering causes. Social scientists assert that specific elements are partial causal factors in producing social reality. Sometimes they have good evidence for doing so.

This seemingly straightforward point has been easy to miss, because defenders such as Cartwright couch it in terms of grand metaphysical theses about capacities, the dappled nature of the universe, and the nature of laws in fundamental physics. But this connection is not essential. The laws of fundamental physics might not be qualified *ceteris paribus* and yet it be true that in physics and elsewhere causal claims are successfully asserted in the face of known and unknown interfering factors.

Likewise, social scientists can assert the existence of causal factors without being committed to capacities, where the latter involve non-occurrent properties. Cart-wright denies this is possible. On her view even the fundamental forces are capacities rather than actually manifested effects—component forces are not real. Cartwright generalizes this implausible view (see Spurrett 2001) to the social sciences and takes them to support her metaphysics. But that metaphysics is not essential to the view that the social sciences try to pick out partial causal factors in the face of unknown and known confounders.

A second standard restriction on causation is that it cannot be about aggregate, social, non-individual entities. This claim is often associated with the idea that any adequate account that has to specify a causal process—a 'mechanism'—in terms of individuals. Despite being widely asserted, on my view these claims are a jumble of ideas most of which are implausible.

Here are some of the things claimed about social causation from this perspective:

> A class cannot be a cause because it is nothing but a constructed aggregate. (Hedstrom and Swedberg 1998: 11)

> We cannot eliminate spurious correlations unless mechanisms in terms of individual behavior are given. (Elster 1983: 59)

> It is not legitimate to assert a causal or explanatory relationship between aggregate social characteristics without at least a stylized conception of how individuals ... bring about these relationships. (Little 1998: p. viii)

> Social entities cannot stand in causal relations because they result from the actions of individual agents. (Varela and Harre 1996)

So there are two fundamental claims being made here: that there cannot be causal relations between aggregate social entities and that such claims are legitimate only if they have individualist mechanisms.

Let's distinguish two different arguments for the first claim. Hedstrom and Swedberg's argument seems to be:

1. Social entities are constructed aggregates.

2. No constructed aggregate can be a real cause.

3. Thus social entities cannot stand in causal relations.

Secondly, there is an argument from supervenience lying behind Verela and Harre's comment and put into general terms by Kim (2005) that seems roughly of the form:

1. Social entities and their properties supervene on the properties of individuals composing them.

2. Any property that supervenes on some more fundamental property has its causal efficacy in virtue of the causal properties of that which it supervenes on.

3. Thus, the properties of social entities are causally irrelevant.

Hedstrom and Swedberg's argument is not compelling, because on various different accounts of causation, it is easy enough to find good reason to think that aggregates can be causes. To see this, it will be helpful to distinguish different kinds of social aggregates: logical aggregates, synthetic aggregates, and complex wholes (Kincaid 1996; Hoover 2001).

Logical aggregates are simple sums of individual characteristics—the total employment is just the conjunction of each individual employed. Logical aggregates are naturally measured in the same units as what they aggregate. Synthetic aggregates are in some sense also constructions, but not ones that are simple sums or averages. The general price level, real GDP, and real interests rates are synthetic aggregates. The dimensions of the general price level are period t dollars/base period dollars while the dimensions for individual commodities is dollars per unit.

Complex wholes are not constructed in either the way logical or synthetic aggregates are. They have internal structure and exist in space and time in a way the other aggregates do not. So a semiconductor firm in the Silicon Valley is a particular with organizational structure.

Using these distinctions, we can see that Hedstrom and Swedberg are both wrong that all social entities are constructed and that constructed entities cannot stand in causal relations.

Take logical aggregates first. If A_1 causes B_1, A_2 causes B_2, and so on, then the totality of As cause the totality of Bs. To see this just plug 'the totality of As' into your favorite account of causality: if the totality of As had not happened, the totality of Bs would not; the totality of As stands in a lawlike regularity to the totality of Bs, etc. Secondly, consider synthetic aggregates. The Fed manipulates interest rates by changing the rate on overnight funds from the Federal Reserve. There may be difficulties in measuring the interest rate or establishing how much influence the Fed has, but we have good reason to think it can manipulate interest rates, we have good reason to think interest rates can stand in causal relations—stronger reasons than we have for believing in Hedstrom and Swedberg's a priori announcement about causation.

Lastly, consider complex wholes. They are not constructed in any sense in the way which logical aggregates are—they are not simply the result of someone's way of classifying or a mathematical result. Of course their existence presumably depends on some individuals being in certain mental states, but not on those of the investigator and are thus real in a way the other aggregates might not be. Moreover, the actions of the particular semiconductor firm have causal consequences on the price and supply of goods; the relevant counterfactuals, regularities, possibilities for manipulation, etc. that mark causation clearly could hold.

Perhaps the supervenience-based arguments formulated by Kim and others build on the intuitions that Hedstrom and Swedberg have. Kim's argument raises numerous issues that I cannot discuss here. The original argument is put in terms of mental entities that supervene on microphysical facts. Focus on purely mental causation also raises issues secondary to the discussion here. But it is clear that Kim intends for his argument to generalize.

We can grant for the sake of the argument that social facts supervene on individual facts in some sense, including the causal facts about individuals. I do not see how that entails that social level causation is otiose. Kim's version of the argument seems more compelling than it should because it puts the issues in terms of properties rather than predicates. Formulated this way it seems that we have two distinct things, the lower-level properties and the higher-level ones, and that the full causal effect is due to the lower-level properties, and therefore the higher-level properties are doing no work. However, there is another way to picture the situation. We can hold that any particular social entity at a given time and its causal powers are token identical with the sum of individuals composing it. That is compatible with social predicates being multiple realizable in individual terms and thus not definable in individualist terms. But that does not commit us to social properties as distinct entities that must somehow relate to individual causation. When a particular corporation acts in a market, it has causal influence. The influence of that specific entity is realized by the actions of the individuals composing it just as the influence of the baseball on the breaking window is realized by the sum of particles composing it. The social level causal claims pick out real causal patterns as types that may not be captured by individual kinds because multiple realizability is real; we may have fully convincing evidence at the aggregate or macrolevel that the causal relation exists without knowing individual level details. But this epistemological difference in how we can identify patterns does not entail a metaphysical difference that must be explained away. Each instance of a social causal pattern is token identical with the causal activity of individuals. The mystery disappears.

Requiring individualist mechanisms for legitimate social level causation seems to me to rest on numerous confusions. Note first that the supervenience argument for the fundamental

role of individuals threatens to degenerate into an argument that causal relations between individuals do not exist either and that any individual explanation must have lower-level mechanisms—neurophysiological? quantum mechanical?—as well. Mechanisms can be described at various levels of compositional detail, so a general demand for mechanisms is no direct argument against social causation unless it is one for individual causal relations as well.

Moreover, 'mechanisms' is a buzzword that can cover quite different things. In particular, mechanisms might be vertical or horizontal. If the Fed influences GDP in raising interest rates because doing the latter increases savings, the savings are a horizontal mechanism, one between the two social entities. This is the notion cited in the quote from Hedstrom and Swedberg. However, the sense of mechanisms invoked by the mere construction and supervenience arguments concern the lower-level entities composing the social causes. These are vertical mechanisms.

Arguments for one kind of mechanism need not be arguments for the other. Thus Elster's claim that we must have individualist mechanisms to prevent spurious correlations makes no sense for vertical mechanisms, for they are not intervening variables. Describing everything individuals did in bringing about the social event of the Fed's raising interest rates will not control for other aggregate variables that confound our evidence that interest rates cause increased savings.

Furthermore, identifying horizontal mechanisms is not necessary to confirm causal claims. The standard clinical trial of a new drug, for instance, usually involves randomization to a control and treatment group with the aim of showing that the drug has an effect; the mechanism of action may not be known at all, but if the trial is conducted correctly we can make well-confirmed causal claims. (And if the thought was that the problem is not one of confirmation but explanation, we explain the outcome without citing the intervening causal process as well.)

Arguably the need for mechanisms and thus for individualist mechanisms cannot be evaluated in the abstract. Not only might mechanisms be horizontal or vertical, they might be described thinly ('some physiological process' in the clinical trial) or thickly (binds antibodies)—we might know much about mechanisms or little, and much about causal processes described at the aggregate level or little. Also, sometimes a macrolevel claim may make strong assumptions about underlying or intervening details and sometimes be much more neutral about them. So, for example, there are macroeconomic theories that require very strong rationality assumptions about individuals and some that make none. There is no reason to think these two kinds of cases stand or fall together. It is reasonable to think that the demand for mechanisms and thus individualist mechanisms must be a function of how much we know and how confident we are in our macrolevel causal claims, in our claims about the mechanisms realizing them, and in our claims that particular mechanisms are presupposed.

So far I have been defending the view that there can be causal relations among social entities and that such claims can sometimes be confirmed without individualist mechanisms. However, in the background of these considerations is an argument for the stronger claim that there must be causal relationships between entities picked out by the social sciences that are not picked out by any other science. The basic intuition here is that there are causal relations that show up at the aggregate level but do not show up at any finer level of detail. Or, in other words, the lower-level details are non-essential. Phenomena of this sort underlie the search for

'universalities' in physics (Batterman 2002), and there is good reason to think that there are similar phenomena in the social realm.

The possibility that social kinds might be multiply realized by the actions of individuals suggests one abstract reason for expecting this to be the case. More concretely, selectionist social mechanisms such as those described in the next section's discussion of functional explanation are strong candidates for causal relations at the aggregate level where low-level details are non-essential. If there is competition between corporations that results in differential survival and birth of certain types of organizations, those causal relations hold so long as the organizational traits are realized, regardless of the kinds of individual incentives, preferences, behaviours, etc. that realize them. Even more concretely, there are compelling examples of aggregate social kinds that seem to stand in causal relations that cannot be cashed out by providing finer and finer individual-level causal detail. Hoover (2001) gives the example of GDP which aggregates over time and which loses its meaning if one tries to specify it in finer and finer detail in terms of individual behaviours. For example, at short enough intervals at the individual level there are times when there is no production going on at all, but the GDP does not drop to zero and then shoot back up a moment later.

These arguments bring up three other issues about causality that the social sciences might shed some light on but which I can only mention here as worth further exploration: causation at a level, simultaneous causation, and the Markov Condition in causal modelling. Hoover argues that the phenomena described above of synthetic aggregates that cannot be decomposed has important implications for the possibility of simultaneous causality and the satisfaction of the causal Markov Condition in causal modelling. The data show that the aggregate macroeconomic variables are correlated even when their past values and values of all other variables are held constant. This suggests there are simultaneous mutual causal relations. The standard response would be to claim that there must be some such variables that can be found by taking increasingly finer cuts of what is being measured. Hoover's example above suggests that is not always the case. Because the aggregate variables cannot be broken down into indefinitely finer detail in terms of time, it is implausible to suppose that mutual causation is really a case of X at t_1 causing Y and then Y at t_2 causing X.

The causal Markov Condition (discussed in Chs. 14 and 23 above) requires that any variable in a causal graph must be conditionally independent of its non-descendants, given the values of its parents. This is a condition that it is required for making inferences from vanishing partial correlations in the Glymour et al. (1987) programme. It presumes that when a past cause does not screen off a correlation between two later effects that there is some omitted variable between them that does. Because the temporal grain of macroeconomic aggregates seems irreducible (as argued above), the prospects for identifying such intervening variables in macroeconomics is dim, rendering the inferences of Glymour et al. invalid.

The third issue that social causation raises concerns the idea of causation acting at a level. This notion is common in informal discussion in the social sciences, and has been fundamental in debates over group selection in biology where group selection is described as selection acting on groups, not individual organisms. Similar talk occurs in the social science literature on evolutionary game theory and cultural group selection (see e.g. Boyd and Richerson 2005). There are Kim-type worries that this notion is metaphysically incoherent. If a cause acts on the properties of the group, presumably those properties are brought about by

the properties of the individuals on which they supervene. So the cause must equally be acting on them. This is the basic idea behind Sterelny's critique of group selection in biology (Sterelny and Griffiths 1999).

A route around this problem is to give up the claim that there is some level where the cause 'really' acts and then to put the issue in terms of whether we can pick out causes with the categories at hand. The arguments above would suggest that we sometimes can only pick out causal relations in aggregate, social terms. That suggests that the sensible content of causation 'acting at a level' is epistemological, not ontological. However, this is just the sketch of a view on some very complex issues.

2. SPECIFIC TYPES OF CAUSAL CLAIMS

As concepts of causation get fleshed out in specific disciplines, it is not surprising that generic causal concepts take second stage to specific types of causes. This certainly has been the case in the social sciences, where there have been longstanding debates on the need for and legitimacy of specific kinds of social causation.

A very common type of claim in the social sciences is that some institution, social practice, etc. exists in order to or because it has some particular effect. Marx claimed that the state existed in order to promote the interests of the ruling class. Durkheim thought that the division of labour existed because it promoted social solidarity. These are teleological or functional explanations. They are widespread and widely criticized as inadequate, largely on the grounds that they cannot be given a coherent causal reading. Many have argued (Hallpike 1986) that the most common justification of these explanations in the social sciences—appeals to biological analogies—are quite misguided.

Some confusions can be cleared up by noting that there are two distinct senses of functional explanation. One sense involves identifying the causal role something plays in a system—its standard effects and their interaction with other elements within a system. These seem to be straightforward causal explanations and not uniquely in need of clarification. The real problem surfaces when it is claimed that a social institution not only has certain effects but exists in order to bring them about, which at first glance seems to posit a mysterious causal process.

One route is to argue that these are legitimate explanations but just not causal explanations. So Cohen (1978) claims they are a species of consequence explanations—explaining something because of the consequences it has. This he treats as a species of nomological-deductive explanation where the major premiss is a law stating that when A has a disposition to produce B, A comes to exist.

There are well-known problems with the general picture of explanation assumed here. There can be deductions from laws that do not explain because the laws are irrelevant: it is a law that men who take birth control pills do not get pregnant. Similarly, showing that A exists when it causes B does not show that A exists in order to produce B—the connection might just be accidental. A second problem with Cohen's account is that it requires too much. Larger corporations may exist because they can take advantage of economies of scale. Yet they do not automatically get larger when doing so would have that effect—lack of foresight, resources,

etc. may prevent it.

A more plausible, causal account that captures at least some uses of functional explanation requires three conditions to show that A exists in order to B (Kincaid 1996; 2006):

1. *A causes B.*

2. *A persists because it causes B.*

3. *A is causally prior to B*, i.e. *B* causes *A*'s persistence only when caused by *A*.

The first claim is straightforwardly causal. The second can be construed so as well. At $t_1 A$ causes B. That fact then causes A to exist at t_2. In short, A causing B causes A's continued existence.

The third requirement serves to distinguish functional explanations from explanations via mutual causality. If A and B interact in a mutually positive reinforcing feedback loop, then A causes B and continues to exist because it does so. Yet the same holds for B vis-à-vis A. Functional explanations do not generally have this symmetry. Thick animal coats exist in order to deal with cold temperatures, but when cold temperatures are present there is no guarantee that thick coats arise and surely even if they do, no reason that would cause the cold to persist.

So on this reading, functional explanations are a unique subset of general causal relations. We can therefore now answer two general complaints about functional explanations in the social sciences: that they make an illicit appeal to biological analogies and that they are tautologous or vacuous.

The most general description of a causal system describes a set of individuals whose values evolve through state space. At this level of description we are told very little: current entities stand in some relation to past ones. Natural selection is inevitably an instance of this system, given that it is a causal system. Functional explanations as causal are also an instance. Every causal system is analogous in being a dynamical system. The point here is that whether one set of causal relations is analogous or disanalogous to another depends on the level of description we are using.

So at the most abstract level it is a trivial truth that functional explanations are indeed analogous to Darwinian evolutionary systems in so far as they are causal systems. They are disanalogous in that social entities have no DNA that replicates. But then the HIV virus has no DNA either (it is an RNA virus). We find analogous processes in DNA and RNA organisms despite the differences because we abstract from the details to identify abstract causal patterns.

So do functional explanations commit us to some illegitimate analogy to natural selection? No, because natural selection explanations are just one realization of the above schema which is thus the more general pattern (Kincaid 1996; Harms 2004). A causing B may result in A's persistence by means that don't involve genetic inheritance, literal copying of identifiable replicators distinct from their vehicles or interactors, etc. In fact not all biological processes of natural selection require this level of analogy—differential survival can be caused by other processes (see Godfrey-Smith 2000).

Another set of specific types of causal claims common in the social sciences involves the

idea of necessary causes, sufficient causes, and important causes. Some quite innovative work has been done on these topics that goes some way towards making these intuitive but vague notions more precise. Ragin (1987; 2000) and others have taken Boolean algebra, used to describe electronic switching systems in engineering among other applications, and have shown how to use it to represent necessary, sufficient, and important causes. Necessary causes are ones always present and sufficient causes are ones that require no other causal factors. The real interest comes not from the obvious definition but from the ability of the Boolean algebra to handle complex configurations of elements to sort out which complexes are necessary or sufficient causes.

At its most elementary, the approach uses nominal variables (presence or absence of a condition) to construct truth tables. Addition and multiplication can then be defined and rules established for simplifying complex configurations into the logically minimum equation describing the different configurations of factors associated with a given outcome. So if A = booming product market, B = threat of sympathy strikes, C = large strike fund, S = strike occurs, and the lower-case versions of these variables represents the absence of these conditions, then a set of data can be used to construct a truth table. Suppose the truth table shows that $S = AbC + aBc + ABc + ABC$, where multiplication is logical 'and' and addition the logical 'or.' The truth table can then be perspicuously described by simplifying its equation to prime implicants—configurations that subsume other configurations. Thus in the above, AC is a prime implicant for ABC and AbC. Logically essential prime implicants are those necessary to cover all the primitive expressions. So

$$S = AbC + aBc + ABc + ABC$$

becomes

$$S = AC + AB + Bc$$

with the essential terms reducing to

$$S = AC + Bc$$

Strikes are caused either by the presence of booming markets and large strike funds, or by threat of sympathy strikes.

Necessary and sufficient causes falls out of this formalism nicely:

$S = AC + Bc$: no cause is either necessary or sufficient

$S = AC + BC$: C is a necessary but not sufficient cause

$A = A + Bc$: A is sufficient but not necessary

One standard doubt about identifying necessary causes is that doing so is trivial, since there are indefinitely many necessary conditions for any effect (Downs 1989). However, there are natural ways to characterize trivial vs. important necessary causes (Goertz 2003). If X_1 is a

necessary cause of *Y*, then *Y* is a subset of X_1. For other necessary conditions X_2 ... X_n, either X_1 is a subset of them or they of it. Necessary conditions are then more important the 'closer' they are to *Y*, i.e. as $X_i - (X_i$ or *Y*) approaches the empty set. From this definition measures for the importance of sufficient conditions can be developed. This is not the only route to defining measures of importance (see Goertz 2003). But the interesting upshot is that there can be intuitive and definite notions of importance that can be brought to debates in the social sciences and elsewhere when there is a desire to talk about significant and insignificant necessary and sufficient causes.

3. ONTOLOGICAL PRESUPPOSITIONS

Causal claims made in the social sciences and the evidence given for them broach interesting issues about what the social world must be like for those claims to be true. These kinds of ontological presuppositions and their epistemological implications are the subject of this section.

Much social research involves causal modelling. The modelling involves either individual equations or sets of equations that describe dependent variables and independent variables with coefficients that intuitively measure the strength of the variable. Those equations are then estimated and tested by statistical measures, usually variants of multiple regression. Regressions are based on data about specified units with various measured characteristics; these instantiate variables and the coefficients identify the estimated average relationship.

A classic case is the literature on inequality which begins with Blau and Duncan's (1978) study of mobility. Using large data sets, they estimate by path analysis the relative influence of years of education, father's status, and other variables on occupational achievement. Subsequent studies added further variables such as intelligence, on the job training, etc.

In their structure these models typically make numerous assumptions about how causality works and how the social world is organized. These assumptions are contingent and contestable claims, illustrating the thesis of the introduction that much of the work in fleshing out causation is domain specific. Some common assumptions are:

1. Fixed entities with attributes: there is a universe of individuals and a fixed set of properties that are distributed among them. So in the work on occupational achievement, it is assumed there is a fixed set of occupations. This precludes occupations coming into and going out of existence, merging, splintering, etc.

2. Constant causal relevance: the causal determinants of occupational achievement do not change over time in Blau and Duncan's models.

3. Common time-frame for causes and effects and for partial causes of the same effect: the measured fluctuations in the causal variables occur in the same time-frame as each other and as the fluctuations in the effect variables. Changes of years of education and changes in occupational attainment are measured over common fixed intervals, precluding the duration of the causing event and the effect event from occurring at different timescales.

4. Uniform effects: the influence of a variable cannot vary according to context. If the

influence of a variable depends on the level of another variable, then the model is misspecified and a further variable representing the interaction effects of the two variables needs to be added. The resultant model then has every variable with a constant effect. In short, context can always be removed.

5. Independent effects: each causal factor makes an independent contribution to the effect—their causal influence is separable.

6. Causal influence is found in variations of mean values.

7. Causation is not asymmetric. Increases in the value of a causal factor will increase the size of the effect and decreases will decrease the size of the effect.

The important point is not that these presuppositions cannot be true. They can be. However, it is also fairly obvious that the complex causal narratives typical of much social science and history describe causal processes that violate some or all of these presuppositions.

The strength of these assumptions can also be seen by looking at alternative methods for representing causes that have been developed that do not make these presuppositions, in particular the Boolean methods of Ragin (1987; 2000) and others mentioned above. Basically, the Boolean approach makes none of the seven assumptions presupposed by standard regression techniques because of its generality. The Boolean approach refers to causal conditions that need not be present, thus allowing basic entities to come and go. 'Causal conditions' is broad enough to allow presence or absence at different time-frames, to allow 'increase in X' to be a separate condition from 'decrease in X', and to allow change in variance, say, as a condition rather than change in mean value. Causation can be conjunctional, such that a condition has causal relevance in one combination and not in another and thus has no uniform or independent effect.

Many social scientists have responded to these complex causal situations by simply assuming they are not the case; many a regression study draws causal conclusions when there is no reason to think the assumptions of simple causation hold. Others face the problem head on, usually by doing case studies and case comparisons. There is vigorous ongoing debate about how to evaluate such work. It has long been thought that such evidence was 'qualitative' and subject to numerous problems not faced by large N studies. I suspect the rhetoric here both underemphasizes the problems of large N studies and overstates the problems of small N evidence. Proof for the former comes from the many assumptions needed to get clear causal information out of correlations (see Woodward 2006). Proof for the latter claim is the subject of the rest of this chapter.

Here are some of the common criticisms of using single or a small number of case studies to infer causation:

1. The degrees of freedom problem: in case studies the number of variables investigated is often greater than the number of data points, so drawing inferences is impossible.

2. It is not possible to rule out chance correlations with the kind of statistical tests that are standard in large N studies.

3. Case studies frequently select on the dependent variable—they make inferences from looking at a set of cases where a given outcome, for example, revolution, occurs and try to infer from its antecedents the causes. However this procedure is inherently biased, because

cases where the outcome does not happen are ignored.

4. Mill's methods are the commonly used tool in comparative case studies, and those methods rely on very strong assumptions that are unlikely to be satisfied. Those assumptions include the presuppositions that there are not multiple causes of the same outcome and that the causal relation is deterministic. Neither assumption is likely to be plausible.

5. Relatedly, the cross-comparisons of case studies are likely to exhibit what Woodward (2002) calls a failure of modularity. A powerful way of showing causality is to show that if it were the case that the variable of interest were manipulated while holding all other causes constant, then the effect variable would change. This requires we have good evidence about changes in the variable alone—changes that do not influence the effect of other variables. However, in case study comparative research, variables change together and we often do not have enough cases for every variable to vary independently. So we cannot make reasonable inferences about causation.

I think these charges are overblown, though there certainly is no consensus on these issues and much work remains to be done in evaluating when, where, and if case study methods are probative. Let me describe some the responses to these concerns and the controversies that still exist around them.

It is not clear that selecting on the dependent variable is always a problem. It is when standard regression techniques are used. Fixing the value of Y in a regression equation $Y = bX + e$ will ensure that the errors are correlated with the outcome and thus provide a biased estimate in the statistical sense. However, standard regression makes no distinction between necessary and other causes. It will find correlations between other variables and the outcome being studied even when there are in fact no cases of the outcome without the necessary cause and will lump everything into the partial cause category. So the defenders of case study work cannot employ the regression analysis as standardly used.

However, for the kinds of evidence that case study work does provide it is not clear that there is any inherent problem with arguing from relevant outcomes only. Surely it happens in biomedical sciences all the time. There I study the people with the disease and trace the causal process from the suspected key event (infection with HIV virus) through its many intermediate causal states to AIDS. I can do this and have good evidence that HIV causes AIDS without looking at those without the disease (and who in fact can have the virus). In some cases this kind of inference can be made by tracing the process in one individual if I have sufficient background knowledge. So clearly the problem is not inherent in causal inferences but perhaps rather inherent in certain kinds of probabilistic evidence.

Defenders of case study research claim that they can sometimes do something similar to what is done in biomedical science. Case studies are forms of process tracing—breaking up large-scale events into a detailed sequence of events where there is a variety of different pieces of evidence that the events are elements in the causal process. For example, in *Analytic Narratives* (Bates et al. 1998) the authors describe the causes of the emergence of specific norms and institutions in Renaissance Italy. Their descriptions use game theory results to help derive predictions and tease out what should and should not be expected in their specific historical case given their hypotheses about causes.

It is, of course, a large and open question just how far these kinds of methods can go and a large task to give a perspicuous account of how they work. Let me describe some of the ideas social scientists have suggested for doing so.

The question of ruling out chance results is an interesting area of research. I should note first that much quantitative social science research overstates its ability to do so. For much of that research it is unclear exactly what grounds claims of statistical probability because there is no obvious chance set-up. Study data are not a randomly selected sample from a well-defined larger population nor the results of random assignment to treatment and control in a defined population. It is rather the observations on all current country economic data, panel or time series. Sometimes this data is claimed to be a sample from a hypothetical population, namely, the various possible realizations of the main causal variables with further influences that are unknown and randomly varying. However, how exactly this story is supposed to work and what kind of evidence would show it plausible has not been well worked out (see Berk 2004).

I should also note that ruling out chance does not really do all the work it is claimed to do in quantitative social science. There it is common to take ruling out the null hypothesis as strong support for the maintained hypothesis. This is not so for multiple reasons. At the most basic level, ruling out the null hypothesis tell us the probability of seeing this data by chance, that is, the $p(D/H = $ sampling error produced the data). But what we want to know is the $p(Hi = $ various other explanations for the data$/D)$. We cannot come to a conclusion about that without the rest of the information summarized in Bayes's theorem. In many cases there is no pretence of providing that information: one causal model is shown consistent with the data and a rejection of the null, but many other models equally compatible with the data are not ruled out.

Moreover, there are ways to rule out chance broadly speaking in case study research. Dion (1998) and Ragin (2000) discuss specific tests that set a presumed level of measurement error and then ask what is the probability of seeing the observed data by sampling error alone. The number of cases needed to reject the chance hypothesis depends on the postulated error rate and the significance level chosen. Calculations show that relatively solid results can be obtained with sample sizes of the sort used in case study comparative work. Thus the issues here are an interesting and lively area of research.

FURTHER READING

Reiss (2007) is a good general survey of the issues concerning mechanisms. Goertz and Starr (2003) is a collection of articles covering a range of issues about necessary causes. A. Abbott (2001) outlines various assumptions about causality underlying regression methods. Kincaid (2006) surveys issues about functional explanation. Issues concerning case studies and causality are surveyed in Mahon and Ruesche-meyer (2003).

REFERENCES

ABBOTT, ANDREW (2001). *Time Matters*. Chicago: University of Chicago Press.
BATES, R., GRIEF, A., LEVI, M., ROSENTHAL, J-L., and WEINGAST, B. (1998). *Analytic*

Narratives. Princeton: Princeton University Press.

BATTERMAN, R. (2002). *The Devil in the Details*. Oxford: Oxford University Press.

BERK, R. (2004). *Regression Analysis: A Constructive Critique*. Thousand Oaks, Calif.: Sage.

BLAU, P., and DUNCAN, O.(1978). *The American Occupational Structure*. New York: Free Press.

BOYD, R., and RICHERSON, P. J.(2005). *The Origin and Evolution of Cultures*. Oxford: Oxford University Press.

CARTWRIGHT, NANCY (1983). *How the Laws of Physics Lie*. New York: Oxford University Press.

COHEN, G. A. (1978). *Karl Marx's Theory of History: A Defence*. Princeton: Princeton University Press.

DION, D. (1998). 'Evidence and Inference in the Comparative Case Study', *Comparative Politics* 30: 127–45.

DOWNS, G. (1989). 'The Rational Deterrence Debate', *World Politics* 41: 225–37.

EARMAN, J., and ROBERTS, J. (1999). 'Ceteris Paribus, There Are No Provisos', *Synthese* 118: 439–78.

— — and SMITH, S. (2002). 'Ceteris Paribus Lost', *Erkenntnis* 57: 281–301.

ELSTER, JON (1983). *Explaining Technical Change*. Cambridge: Cambridge University Press.

GLYMOUR, C. (2002). 'A Semantics and Methodology for Ceteris Paribus', *Erkenntnis* 97: 395–405.

—— SCHEINES, R., SPIRTES, P., and KELLEY, K. (1987). *Discovering Causal Structure*. New York: Academic Press.

GODFREY-SMITH, P. (2000). 'The Replicator in Retrospect', *Biology and Philosophy* 15: 403–23.

GOERTZ, G. (2003). 'Cause, Correlation and Necessary Conditions', in Goertz and Starr 2003: 47–65.

—— and STARR, H. (2003). *Necessary Conditions*. Lanham, Md.: Rowman & Littlefield.

HALLPIKE, C. (1986). *Principles of Social Evolution*. New York: Oxford University Press.

HARMS, W. (2004). *Information and Meaning in Evolutionary Processes*. Cambridge: Cambridge University Press.

HAUSMAN, D. (1992). *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.

HEDSTROM, P., and SWEDBERG, R. (1998). *Social Mechanisms*. Cambridge: Cambridge University Press.

HOOVER, K. (2001). *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

KIM, J. (2005). *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

KINCAID, H. (1996). *Philosophical Foundations of the Social Sciences*. Cambridge: Cambridge University Press.

—— (2004). 'Are There Laws in the Social Sciences? Yes', in C. Hitchcock, *Contemporary Debates in Philosophy of Science*. Oxford: Blackwell, 68–187.

—— (2006). 'Functional Explanation and Evolutionary Social Science', in M. Risjord and S. Turner, *Philosophy of Anthropology and Sociology*. Amsterdam: Elsevier.

LITTLE, D. (1998). *Microfoundations, Method and Causation*. New Brunswick, NJ: Transaction.

MAHONEY, J., and RUESCHEMEYER, D. (2003). *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press.

PIETROSKI, P., and REY, G. (1995). 'When Other Things Aren't Equal: Saving Ceteris Paribus Laws from Vacuity', *British Journal for the Philosophy of Science* 6: 81–110.

RAGIN, C. (1987). *The Comparative Method*. Chicago: University of Chicago Press.

—— (2000). *Fuzzy-Set Social Science*. Chicago: University of Chicago Press.

REISS, J. (2007). 'Do We Need Mechanisms in the Social Sciences?' *Philosophy of the Social Sciences* 37: 163–84.

SPURRETT, D. (2001). 'Cartwright on Laws and Composition', *Studies in the Philosophy of Science* 15: 253–68.

STERELNY, K., and GRIFFITHS, P. (1999). *Sex and Death*. Chicago: University of Chicago Press.

VARELA, CHARLES R., and HARRE, ROM (1996). 'Conflicting Varieties of Realism: Causal Powers and the Problems of Social Structure', *Journal for the Theory of Social Behaviour* 26: 313–25.

WOODWARD, J. (2002). 'There is No Such Thing as a Ceteris Paribus Law', *Erkenntnis* 57: 303–28.

—— (2006). 'Causal Models in the Social Sciences', in M. Risjord and S. Turner (eds.), *Philosophy of Anthropology and Sociology*. Amsterdam: Elsevier, 157–212.

CHAPTER 37
CAUSATION IN THE LAW

JANE STAPLETON

1. OVERVIEW

Previous accounts of 'causation' in the law are flawed by their failure to appreciate that causal language is used to express different information about the world. Because causal terms have been used to communicate answers to different questions, any philosophical search for a free-standing account of causation is doomed. Lawyers require precision of terminology, so they should explicitly choose just one interrogation to underlie causal usage in law. In my radical account I argue that this interrogation should be chosen to serve the wide projects of the law. In these projects the law is interested to identify when a specified factor was 'involved' in the existence of a particular phenomenon, where the notion of 'involvement' identifies a contrast between the actual world and some specified hypothetical world from which we exclude (at least) that specified factor: this contrast being that, while in the former world the phenomenon exists, in the latter it does not. (Such contrasts of necessity can be generated in three ways, all of importance to the law.) Because this determination of involvement is made using our knowledge of the physical laws of nature, evidence of behaviour, and so on, choosing 'involvement' as the meaning of 'causation' in law carries the potential for this concept to be untainted by normative controversies. This is the most convenient choice for the law because (a) it promotes clarity and avoids ambiguity; (b) it promotes the clear identification of normative issues and provides a more transparent distribution of issues between causation and other analytical elements within legal analysis; and (c) it best serves the law's very wide range of purposes.

2. INVOLVEMENT

The world is out there, seamless and rolling along manifesting what we call the physical laws of nature in complex confluence and combinations. Just as we use a variety of limited interrogations to investigate this complex world and discover its underlying building blocks of

physical laws, so too we can investigate with a variety of limited interrogations a particular phenomenon, be it the persistence of a specified state (e.g. the cannonball resting on Kant's cushion) or a transition from a specified state to another (e.g. the creation of a statue, of which Aristotle suggests we might ask: Who made it? From what raw materials? To what pattern? And for what purpose?)

Often our interrogations focus on a specified factor—such as a physical force, an absent factor or the communication of a specific piece of information—and its relation to the existence of a specified phenomenon. Different interrogations yield different, limited sorts of information about such relations in the world. For example, take the case where, due to the carelessness of each of two unrelated hunters, a mountain walker is simultaneously shot by both and the medical evidence is clear that either shot would have been sufficient to result in instantaneous death; and a hunting official who had promised the hunters that he would shout a warning about the presence of any walkers had remained silent even though he had seen the mountain walker clearly.

One legal interrogator might ask: do we blame Hunter No. 1 even though the death would have happened anyway? (yes). A citizen might ask: did the victim's presence explain his own death? (no). A scientist might ask: did gravity play a physical role in the death? (yes). Sometimes interrogations explicitly compare what actually happened with some designated hypothetical world such as: in order to prevent the death, would the carelessness of Hunter No. 1 have to have been absent? (yes); did the carelessness of Hunter No. 1 make a difference to the occurrence of the death? (no, in a hypothetical world without that conduct the death would have happened anyway); or did the official's breach of promise, an omission, make a difference to the occurrence of the death? (yes, in a hypothetical world in which he communicated a warning to the hunters, the death would have been prevented).

What does the diversity of such interrogations and their results have to do with 'causation'? The answer is that the different types of information resulting from these different interrogations are often expressed in the same causal terms. For example, the citizen might express his conclusion from his interrogation (namely, into explanation) in terms of the victim's presence not being a cause of his death while a scientist might express the conclusion of his investigation (namely, into factors that played a physical role in the death) in terms of the victim's presence being a cause. Because the same causal language has been used to convey different types of information, it is futile for lawyers or philosophers to search for a coherent free–standing account of causation. Unless a choice of underlying interrogation (blame, explanation, physical role, any sort of involvement, etc.) has been specified at the outset, we simply cannot say whether to be successful an account of causation must identify the mountain walker's presence as a cause of his death, or identify it as 'not a cause' of it.

In law terminology should be as clear and unambiguous as possible, so lawyers should explicitly choose just one interrogation to underlie causal usage in law so that it is clear what information we are reporting when we use causal language for legal purposes. The most appropriate choice of interrogation to underlie causal terms in law is not self-evident and could be influenced by a number of concerns on which reasonable minds might differ. Nevertheless, I argue (Stapleton 2008) that this interrogation should be chosen both to serve the wide projects of the law and to be such that it is untainted by normative interrogations and

controversies.

From a consideration of the various projects of the law we can construct what must be accommodated within this interrogation to underlie causal usage. For example, the interrogation must be capable of accepting an *omission* as the 'specified factor': the law is often engaged in investigating the significance of omissions in the existence of a particular phenomenon, such as when a child dies following a parent's neglect or when a lifeguard, in breach of his legal obligation, fails to try to rescue a toddler drowning in the shallows of the surf. Similarly, the interrogation must be capable of accepting the *communication of a specific piece of information* as the specified factor, because the law is clearly concerned with the significance of such factors, as when a fraudster dishonestly gives a bank false information in the hope the bank will lend him money in reliance on that information.

In contrast, the law is often concerned with the significance of *innocent and/or normal* factors in the existence of a phenomenon. Again, the chosen interrogation must accommodate such factors as the specified factor of interest. For example, in the hunters' case one project of the law might be to consider all possible regulatory strategies for preventing such deaths, and so it needs to address all factors relevant to the occurrence of these deaths—even those that some might describe as 'mere conditions', such as the walker's presence—because the most efficient strategy may be to ban mountain walking during the hunting season. Were lawyers to select, as the interrogation underlying causal language in law, one that only identified factors that were blameworthy, abnormal, or explanatory, this regulatory project might be inhibited.

As to the form of a specified factor's 'significance' in the existence of a particular phenomenon, this last example illustrates how the law is often interested to identify when, although the particular phenomenon existed in the actual world, it would not have existed in the hypothetical world notionally constructed by simply removing the specified factor from the actual world. We can convey this information by saying that the factor was 'involved' in the existence of the particular phenomenon by being 'necessary' for it according to the data of our natural world: for example, the walker's presence was necessary for his death (on this form of necessity, see Stapleton 2008: 438).

But the law is also interested to identify another relation between the specified factor and the existence of the phenomenon: when, although the particular phenomenon existed in the actual world, it would not have existed in a hypothetical world notionally constructed by removing the specified factor from the actual world along with any factor that duplicated the relevant effect of the specified factor. We can convey this information by saying that the specified factor was 'involved' in the existence of the phenomenon by a relation of 'duplicate necessity': for example, the careless conduct of Hunter No. 1 was involved in the death of the walker by a relation of duplicate necessity.

A third relation of interest to the law can be illustrated thus: nine members of a club's governing committee unanimously vote in favour of a motion to expel Member *X* from the club, where a majority of only six was needed. The vote of Committee Member No. 1 is neither necessary nor sufficient for the motion to pass. This is true of the vote of each member, yet the motion passed. Each voter must have played some role in the passage of the motion: a role I call 'contribution' to the existence of that particular phenomenon. The law is concerned to identify the contribution of individuals such as Committee Member No. 1. This can be done by considering a subset of factors identical to the actual world except it lacks the

votes of Committee Members Nos. 7, 8, and 9 (in other words, factors in excess of what is sufficient for the motion to pass). In this subset the motion would have passed but only because of the presence of the vote of Committee Member No. 1. We can express this information by saying that the vote of Committee Member No. 1 was 'involved' in the motion passing by 'contributing' to it.

Finally, the interrogation underlying causal usage in law must be capable of conveying information that contrasts the actual world and a hypothetical world *even where the latter was impossible*. This is because the law often specifies as its factor of interest a defendant's breach of legal obligation. Here the notion of 'involvement' compares the actual world with a hypothetical world in which,

inter alia, that breach is absent and the obligation is fulfilled. Lawyers fully appreciate how artificial such hypothetical worlds are: indeed, it may have been impossible for the defendant to have fulfilled his obligation. An example of such a hypothetical world is one in which a retailer would have made a $30 profit had a farmer fulfilled his contractual promise to the retailer to deliver peas on a certain day ... a delivery which turned out to be impossible because of the carelessness of a third party. It is important to understand how it is that, in the context of that comparison, lawyers convey meaningful information when they say that the farmer's contractual breach (his omission to deliver the peas) 'caused' the retailer to suffer a $30 'loss': see below.

In my view lawyers should choose as the interrogation underlying causal usage in law one that captures 'involvement' in any of these forms. That is, they should choose an interrogation that accepts that the relevant specified factor may be an omission, a normal and/or innocent factor, or the communication of a specific piece of information; and that the hypotheticals in play are artificially constructed from the removal, *inter alia*, of that factor. This would, for example, identify the vote of Committee Member No. 1 as a cause of the motion of expulsion passing, even though that vote was neither necessary nor sufficient for the existence of that particular phenomenon.

The choice of having causal terminology in law refer to this broad relation of 'involvement' presents no problem of over-inclusiveness (that is, too many causes) because doctrinal filters ensure the legal enquiry is tightly focused: see below. Moreover, the involvement choice has positive advantages over alternative approaches to what should be denoted by causal language in law. First and obviously, it can provide the width of coverage to accommodate smoothly all the many diverse enquiries law makes because it was from a consideration of these needs that the interrogation was chosen. For example, the involvement interrogation captures the involvement of a factor in an outcome even where that outcome is a coincidental consequence of the factor (that is, where factors of that type do not generally increase the probability of outcomes of the designated type): this is a relation with which the law can be concerned, as when the dishonesty of a first wrongdoer results in the money of the victim being invested in a company, the share price of which later collapses when the market discovers the unrelated fraud of a company official.

Secondly, because 'involvement' is identified by objective data (our knowledge of the physical laws of nature, evidence of behaviour, and so on), the concept of causation in law will

be untainted by normative interrogations and controversies (though, of course, the rules of *proof* of causation and their exceptions are normative determinations). These must therefore be located elsewhere in the legal analysis where traditionally they are more likely to be exposed as normative issues and evaluated accordingly.

For example, suppose *A* inadvertently discloses to *X* a fact about *V* that enrages *X*, *B* encourages *X* to kill *V*, *C* knows *X* is planning to kill *V* and does nothing to stop *X*, *D* suspects *X* wants to kill *V* but gives him a gun anyway, *E* threatens to kill *X* unless *X* kills *V*, *X* shoots *V* dead. *A, B, C, D, E,* and *X* are all 'involved' in the death of *V*. By choosing an interrogation underlying causal usage in law that identifies but does not distinguish between different forms of involvement, we allow each of these parties to be identified as a cause of the death. This then requires us to locate the normative controversies about their different degrees of responsibility for the death under analytical labels such as 'duty', 'breach', 'aiding and abetting', 'complicity', 'inducement', 'duress', 'solicitation', and so on. The great attraction of this approach is that under these analytical labels, unlike the label of 'causation', it has traditionally been unacceptable merely to assert a conclusion on the basis of 'intuition' or 'common sense'.

Finally, once we have chosen 'involvement' as the question underlying causal usage in law, we can then test for it using the NESS algorithm originating in the work of Hart and Honoré (1959).

3. PHILOSOPHICAL ISSUES THE LAW CAN IGNORE

Traditionally, most lawyers disdain philosophical enquiries into causation. In my opinion this indifference is warranted. Many philosophers do not seem to agree on which underlying interrogation their causal expressions refer to. Some seem to use the 'intuition' of 'folk' (or more correctly non-empirical assertions of linguistic usage) as the benchmark against which a philosophical account of causation is assessed, even though usage is contingent on time and place. Others seek to bathe their account of causation in scientific respectability by reference to a crude push–pull concept of physics, one that ignores the role that comparison and absences play in scientific understandings. Yet, as I argued above, if causal language is used to express the results of quite different interrogations of the world, it will never be possible to formulate a reductive algorithm that will detect when some factor is, metaphysically, a cause.

In any case, by its very nature the social practice of the law can afford to ignore many issues that philosophers understandably find problematic. Just as in everyday life, in the law we can accept (though we may be unable to account for it) that time is not reversible; that there are lawful regularities manifesting fundamental laws of nature (which are to be distinguished from the law-like regularities of a mere association, a mere 'constant conjunction', such as the epiphenomenal fall of a barometer before a storm); that 'God does not play dice', so that given a sufficiently detailed description of initial conditions within a closed system, the state of that system at a later time may be calculated according to fixed laws of nature; and that, though the world is deterministic, proof of its phenomena may have to resort to probabilistic evidence. Similarly, the epistemic concerns of philosophers are subsumed within the law's acceptance of

expert scientific evidence (even when based on induction), witness statements and so on.

Moreover, in contrast to the context-free project of philosophers (e.g. Schaffer 2005), the conceptual framework and methodology of the legal project provide contextualizing devices that render finite the number of factors whose possible involvement is subject to investigation; that individuate (at a level of modal fragility to serve law's purposes) the specified factor whose influence is being examined and therefore the hypothetical worlds with which comparison is being made; and that individuate the phenomenon (again, at a level of specificity to serve law's purposes) in relation to which that influence is judged, an individuation of outcome that often excludes any problem of pre-emption. Because doctrine focuses the lawyer's project in this way, just as scientific method focuses scientific interrogations of the world, the meaning of involvement can be sufficiently unambiguous. This means that, once lawyers choose involvement to be the underlying interrogation they can ensure that causal language in law conveys a sufficiently unambiguous meaning.

The central importance of these doctrinal filtering devices may be dangerously overlooked by non-lawyers, so I will illustrate their operation using the area of law on which most theoretical discussions of 'causation in the law' focus: the tort of negligence.[1] This is a species of legal claim, a 'cause of action', which a plaintiff may initiate against a defendant.

Suppose a parent, *D*, fails to control their 2-year-old infant who runs out into the path of a moving vehicle which swerves and breaks the leg of a pedestrian, *P*. On the way to hospital, the ambulance carrying *P* is struck by lightning and *P* is seriously burnt. *P* sues *D* in the tort of negligence for his broken leg and burns. To succeed in the claim *P* must prove, on the balance of probabilities, all five doctrinal elements of this cause of action.

The first three elements are: that the interference of which *P* complains is recognized as a form of actionable damage; that *D* owed *P* a duty of care; and that some aspect of *D*'s conduct was a breach of that duty because it fell below the standard of care that a reasonable person would have met in the circumstances. The fourth element is the causation requirement: that this breach was a cause of the injury of which *P* complains. The terminology of the fifth element varies: in the US the requirement is generally said to be that the breach was a 'proximate cause' of the injury; while in the non-US common-law world the requirement is said to be that the injury must not be 'too remote' a consequence of the breach. Functionally, this fifth element sets the normatively appropriate scope of liability for consequences of breach by *D*: only certain of the infinite stream of consequences of breach will be judged to be appropriately attributed to the defendant's breach for the purposes of legal liability. There is a strong modern move to rename this normative determination concerning for which consequences a defendant should be legally responsible, as the 'scope of liability' issue (Stapleton 2001*b*; 2003). In this chapter I will refer to it as the 'proximate cause/scope' issue.

Taking the elements in turn: the law specifies (that is, it is a question of law to be decided by judges) whether the type of injury of which *P* complains is actionable damage in the relevant species of legal claim. Whether a duty is owed is also a question of law. For normative reasons the law does not often impose a duty of affirmative action to control the conduct of others, so in our example this duty element provides a very effective filter. The fact that there were an infinite number of people who failed to control the infant is not a problem since the law will have imposed a duty to control the infant on only a tiny number of them, such as the infant's parents. Claims against others will fail at this early stage in the legal

analysis.

Next, the plaintiff must specify the aspect of the defendant's conduct that he claims fell below the standard of reasonable care: this is termed 'the tortious aspect' of *D*'s conduct or *D*'s 'breach' of the standard of reasonable care. Much rides on this formulation. First, the judge or jury may not accept that a reasonable person would have behaved as the plaintiff alleges the defendant should have behaved.

Secondly, and of more importance here, the plaintiff's individuation of the breach allegation can affect plaintiff's chances of establishing, as he must, that the breach qualified as a cause of the relevant injury. This is because that individuation, if accepted by the law as reflecting what a reasonable person would have done, constitutes what I have been calling the 'specified factor'; and the specified factor in turn determines the hypothetical worlds (because these are no-breach worlds) against which the law is concerned to identify any involvement of that factor in the existence of the actual phenomenon (Stapleton 2008).

A useful illustration of how the law chooses the relevant hypothetical worlds can be drawn from the law of contract: say on Sunday a farmer sells peas to a retailer for the current wholesale price of $100 a bushel, stating that they will be delivered on the following Tuesday. Since the current retail price is $150, the retailer thinks he has a good bargain. Though the farmer exercises all reasonable care, he is prevented from delivering the peas until the following Friday by the carelessness of a third party. On the Tuesday the retail price of peas has fallen to $130 per bushel and by the Friday when the retailer sells the load of peas delivered by the farmer it had fallen to $100. Under the law of contract the retailer can recover compensation for any loss that was 'caused' to him by any breach of contract by the farmer. To see whether our retailer can recover compensation the court would need to decide what sort of performance from the farmer the retailer was entitled to under the contract.

If the farmer's statement is determined by the court to be merely a promise that care will be taken to deliver on time, an allegation of breach of contract will fail because the farmer took care: there is no difference between what happened and a hypothetical 'no-breach world'. If, however, the statement is held to be a contractual promise as to result (namely, that the peas would be delivered on Tuesday), there is a breach of contract constituted by the farmer's failure to deliver on Tuesday. Now there is a difference between what happened and a hypothetical 'no-breach world' in which the peas were delivered on Tuesday. In the context of that comparison we can see that the breach was involved by a connection of necessity to the phenomenon of interest, namely the retailer's static financial position, which represents a 'loss' relative to where the law says he was entitled to be. The quantum of that loss is $130 (what he would have recouped in a no-breach world) minus $100 (what he actually recouped) = $30, which he can recover from the farmer. Lawyers would express this as the farmer's contractual breach having 'caused' the retailer a $30 loss.

That the law requires the precise specification of its investigation—including the precise specification of the factor in issue (for example, the precise breach allegation) and the particular phenomenon in whose existence that specified factor may have been involved in some way—explains why the law does not perceive problems where philosophers do. For example, unlike many philosophers (e.g. Hall 2004: 241–8), the law sees no problematic 'double prevention' problem where the specified factor prevented another factor from preventing a specified outcome. Suppose that, as a result of the defendant's wrongful conduct,

either by act or omission, there is no stop sign at a road crossing. This absence prevents a careful motorist being alerted to the need to stop and thereby avoid a collision with another driver, *V*. That collision occurs and knocks *V* unconscious. To determine whether the defendant's wrongful conduct was involved (in the sense of necessity) in *V* being rendered unconsciousness, the law considers what *V*'s fate would have been in the absence of the wrongful conduct: that is, where the defendant had conducted himself lawfully and as a result a stop sign was present. With a stop sign present, the driver would probably have stopped and the injury to *V* would have been avoided. Once this context is specified lawyers can coherently express this information as the defendant's wrongful conduct being involved in, that is a 'cause' of, *V* being rendered unconscious.

Next, consider the individuation of relata. The law distinguishes, for clear normative reasons, the role of distinct actors and this affects the acceptable individualization of the specific factor whose involvement is at issue. Take our earlier case where a mountain walker is simultaneously shot by two unrelated hunters and the medical evidence is clear that either shot would have been sufficient to result in instantaneous death. Since the law is concerned, *inter alia,* with the role of individual defendants (*pace* Mackie 1974), it focuses not on whether the 'cluster' of the conduct of the two hunters was involved in the relevant outcome but whether a specified hunter was, treated as an individual, involved in that outcome.

Similarly, law has bases on which it individuates the outcome of which complaint can be made. A well-known illustration is as follows: a boy fell from a high bridge spanning a vast canyon but on the way down he hit and was fatally electrocuted by power lines strung carelessly and in breach of duty by a power company. When his estate claimed against the company, the law required that the estate specify the outcome precisely as 'death by electrocution at that instant' rather than, say, 'death on that day'. The fact that his death was certain to have happened seconds later by a lethal back-up phenomenon (crushing impact with the canyon floor) will therefore be irrelevant to the particular phenomenon of interest to the law (namely, the actual death of the boy by electrocution at that instant) and will therefore be irrelevant to the question whether the company was a 'cause' of and liable for the 'death' as individuated in this way. (The boy's prospects are, of course, relevant to how much the company will have to pay in damages.)

Our hunters' case also illustrates how the law individuates the particular phenomenon: for in that case the law focuses on whether the specified hunter was involved in an outcome framed in terms of the walker's 'death by gunshot at the relevant time and place' rather than an outcome whose delineation is as finely grained as 'death by two bullets at the relevant time and place'.

4. Other Theoretical Accounts of 'Causation in the Law'

Having sketched the case for law to choose involvement as the sole interrogation underlying causal language in law, in this section I set out other theoretical accounts of causation in the law. These are inferior for a variety of reasons. For example, the approach of early US Realists such as Leon Green could not satisfactorily identify duplicate involvement while the linguistic analysis of Hart and Honoré confusingly conflated the objective issue of

involvement with the normative proximate cause/scope issue of whether a factor's involvement in the specified outcome was sufficiently relevant to the purpose of the enquiry for legal consequences to flow.

4.1 Early US Realists: A 'Minimal' Factual Concept

The development of significant modern accounts of causation in law begins with the American Realists, especially Leon Green (Robertson 1978). From the 1920s, American Realists argued that many legal rules and concepts were 'over-general and outworn abstractions' (Oliphant 1928: 75). Such concepts were 'rationally indeterminate' because they were subject to 'equally legitimate, but conflicting, canons of interpretation' (Leiter 2005: 64) and as such they could not justify a unique decision by an ultimate court of appeal which was not bound by precedent. Moreover, they often obscured the true reasons for appellate judgment (Leiter 1999: 1147) which were primarily stimulated by the facts of the case and non-legal considerations rather than by legal rules and concepts. Realists argued that legal analysis should be refashioned to expose the judges' real reasons for decision.

Green (1927) showed that the ambiguous use of causal terminology was one technique by which such obfuscation was greatly facilitated: in cases where the defendant's wrongdoing was clearly involved in producing the plaintiff's injury, courts were rejecting plaintiffs' claims by merely asserting that the wrongdoing was not a 'proximate cause' of the injury or that the 'chain of causation' had been broken or that the injury was not a 'natural consequence' of the wrong-doing. In other words, causal terminology was used by courts not only in relation to the involvement enquiry but also to communicate findings about the normative proximate cause/scope issue. For Realists such as Green these proximate causation devices were 'word magic whereby unprincipled limitation-of-liability decisions could be achieved at will or whim by untrammelled judges' (Robertson 1996–7b: 1114). The deployment of causal language here falsely suggested some scientific rationale for these normative decisions and obscured their real basis: namely, the determination that the relevant consequence of the wrongdoing fell outside the scope of liability that was judged appropriate for the particular legal rule in the light of its purpose.

Green argued in favour of separating out the factual enquiry of involvement from normative considerations. The potential strength of Green's approach was that by stripping out such considerations and locating them elsewhere in the legal analysis of the case, they could be identified and evaluated for their normative soundness. Indeed, so attractive was this idea of separating out the factual enquiry that it is now the orthodox position in the US private law (Dobbs 2000: 409).

But there were difficulties in Green's theory. First, in his later writings (Green 1972; 1962: 546) Green argued that the causation enquiry should precede the analytical filters of duty (into which he folded case law on proximate cause) and breach. This arrangement in turn necessitated that the subject of that 'gateway' enquiry into causation should be the defendant's conduct as a whole and not its 'tortious aspect'. This odd analytical arrangement proved too awkward in application to be persuasive. For example, in theory it would inefficiently require

the causal analysis to be applied to the vast number of omissions that may have been involved but were legally of no concern because they were lawful and not breaches of a recognized legal duty.

Secondly, Green's approach was fatally incomplete in important respects. For example, Green was unable to identify a coherent test for involvement. He rejected *sine qua non* as an adequate test because, *inter alia*, it failed to identify duplicate involvement leading to an overdetermined outcome such as the conduct of each hunter in the earlier example about the two hunters (Green 1928–9: 604–5). However, Green (1927: 137) and contemporaries such as Becht and Miller (1961) were unable to formulate an algorithm for involvement that adequately encompassed duplicate involvement and were forced to resort to the very sort of obfuscatory slogans that Realists generally deplored: that a factor qualified as a 'factual cause' if it was a 'substantial factor' in bringing about the outcome. Even later scholars who were broadly admiring of Green astutely confined the role of the substantial factor device to that of an adjunct to the but-for test, where resort to that adjunct was only to be tolerated in overdetermined outcome cases (e.g. Robertson 1996–7*a*: 1776–8).

Thirdly, even though most US courts and commentators followed Green and cleanly distinguished the involvement issue (described as 'factual causation') from the normative proximate cause/scope issue, the latter did not shed its causal terminology, typically being described, at least until recently, as the proximate cause issue. Most modern courts acknowledge that the issue has nothing to do either with factual cause or proximity (Robertson 1997: 10). The risk that this odd legal usage would mislead juries and others has generated concern (e.g. Stapleton 2001*b*; 2003), prompting the American Law Institute to reformulate the law on the issue in the non-causal language of the 'scope of liability for consequences of breach' (American Law Institute 2005: 574–5), banishing the term 'proximate cause' and confining all causal terminology to the analytically prior issue of involvement which is termed 'factual causation'. A law reform body in Australia (Commonwealth of Australia 2002: 109), drawing on Stapleton's critique, recommended the same bifurcation of issues in private law into 'factual causation' and 'scope of liability': this has now been adopted in legislation in all Australian jurisdictions.

Finally, Green often gave the impression that outside the factual cause area the courts were faced with an amorphous mass of normative concerns that had not been, and perhaps could not be, structured. This exaggerated 'scepticism as to the possibility of framing rules' (Hart and Honoré 1959: 92) relating to the attributions of responsibility for consequences provoked a landmark work by Hart and Honoré.

4.2 Linguistic Analysis

In 1959, Herbert Hart and Tony Honoré published a major study of 'causation in the law' using the then fashionable tools of linguistic analysis. Hart and Honoré asserted that 'it is the plain man's notions of causation (and not the philosopher's or scientist's) with which the law is concerned' (1959: 1) and should be concerned. They believed that these non-legal 'common-sense' notions of causation 'have very deep roots in all our thinking and in common ideas of when it is just or fair to punish or exact compensation' (ibid.). In other words, what Hart and Honoré meant by 'the concept of causation in law' seems to be merely an artefact

derived from how lawyers used causal words, which they asserted was the same way that ordinary people used them. Theirs is not a metaphysical account (Lipton 1992: 130; Hancock 1961: 150).

They examined how causal language was used in the two traditional analytical steps of 'cause' and 'proximate cause'/'remoteness' (now 'scope') and found four distinct notions. One was the idea of a 'causally relevant condition' (Hart and Honoré 1959: 107) which is roughly equivalent to the relation I have called 'involvement'. For this notion Hart and Honoré provided an elegant and largely successful algorithm which has later become known as the 'NESS' test, discussed below (ibid. 104–8, 116–19, 216–29).

The second notion emerged, according to Hart and Honoré, in explanatory enquiries where they noted that ordinary language did not characterize every causally relevant (involved) factor as a cause of a contingency 'the occurrence of which is puzzling because it is a departure from the normal, ordinary or reasonably expected course of events'. (ibid. 31) According to Hart and Honoré (1985: 72), in explanatory enquiries the term 'a cause' is reserved for a factor that is an abnormal feature of the situation or a free deliberate human action: an intervention on a 'stage' already set. Thus they noted, for example, that an omission to take a normal precaution could be a cause (1959: 35).

As Foot (1963: 507) points out, the objection to Hart and Honoré's treatment of explanatory enquiries is that their proposition, that a factor which explains a 'departure from [the] normal ... course of events' is an 'abnormal' feature, suffers from circularity: 'the form of the question determines the form of the answer'. Moreover, in the context of many legal enquiries the law is concerned with departures not from normality, but from the mandated course of events. As we saw earlier, the law identifies what was mandated *before* it investigates whether the defendant's alleged departure from that obligation was a cause of the outcome. Departures from the mandated course of events may be normal practice among those in the position of the defendant and yet adherence to such 'normal practice' can attract liability when it results in harm. Courts need to, and typically do, describe such a breach of obligation as a 'cause' of the harm.

According to Hart and Honoré a third notion, that of 'causal connection', could be identified from an analysis of how language was used in attributive enquiries such as the proximate cause/scope stage of legal analysis. Hart and Honoré asserted that even if a factor was 'causally relevant' and a 'cause' in an explanatory sense (i.e. an abnormal feature of the situation or a free deliberate human action) more was needed before a legally adequate form of causal connection could be established between the factor and the specified outcome. In other words, Hart and Honoré (1959: 123) believed they were able to expose 'a group of causal notions embedded in common sense' that influenced when legal responsibility was truncated, thereby refuting the assertion by Leon Green that proximate cause/scope was an area that could not be expressed in clear principles.

Hart and Honoré distilled these notions of causal connection (which they characterized as both causal and factual: 1985: pp. lii, 91) from observations of how courts respond to interventions occurring after the defendant's breach and before the injury to the plaintiff. The first and 'central' notion of causal connection is in operation, say Hart and Honoré (1959: 123), where courts cite intervening 'voluntary action [by another] or abnormal and

coincidental events as negativing causal connection'. For example, suppose that while speeding a motorist loses control, hits a pedestrian and breaks her leg. On the way to hospital the ambulance carrying the pedestrian is struck by lightning and she is killed. It is virtually certain that the driver will not be held legally responsible for the death. Under the terminology used by US lawyers at the time Hart and Honoré were writing, the plaintiff would have been unable to establish that the breach was the proximate cause of the death. According to Hart and Honoré (1985: p. xlv), 'causal connection in the ordinary [central] sense' was negatived by the lightning.

But sometimes a careless party *is* held liable for injuries resulting from the intervention of, say, lightning or voluntary human conduct. These cases led Hart and Honoré to identify a second form of causal connection that was not severed by the intervention: namely that of 'occasioning harm', for example 'by providing opportunities' for such an intervention (1985: 59, 194–5). Finally, Hart and Honoré were confronted with cases where the defendant was held liable for inducing an intervention by another. An example of this is where a fraudster induced the plaintiff to enter a transaction that resulted in loss to the plaintiff and the fraudster was held responsible for that loss. Hart and Honoré responded to such cases by identifying a third form of causal connection that was not severed by the other changing his position in response to the defendant: namely, that of 'providing reasons' for the other to change his position.

Truncation of legal responsibility by these notions of causal connection had, Hart and Honoré argued, been unhelpfully amalgamated, under the common term of 'proximate cause' or 'remoteness', with a fourth distinct set of ideas, namely 'non-causal' limitations on the appropriate scope of liability. An example of this fourth type of concern was when liability for a consequence of the defendant's wrongdoing was excluded on the basis of 'the optimum allocation of social risks ... [and] the impact in a given case of the equities as between the parties' (1985: p. xxxvi).

The approach of Hart and Honoré is inferior to the involvement approach for a number of reasons. First, as we have seen, their approach has difficulty accommodating the fact that the law needs to and does identify normal departures from a mandated standard as 'causes'.

Secondly, Hart and Honoré neglected the critical role played by the formulation of the alleged breach in the attributive proximate cause/scope enquiry. For example, while the law of negligence refuses to impose legal responsibility for the lightning injuries in the ambulance example, liability has been imposed where defendants carelessly 'allowed inflammable vapour to remain in the bottom of a barge ... [and it] was ignited by a flash of lightning' (Hart and Honoré 1959: 184). Exactly the same type of intervention has occurred but liability is sometimes imposed and sometimes not. The approach of Hart and Honoré obscures the fact that the differential lies in the relation of the breach to the consequence, see below.

Similarly, the nature of the breach of obligation, neglected by Hart and Honoré, provides a more coherent explanation of why in some cases but not others the law refuses to impose legal responsibility on a defendant when the voluntary act of another person has intervened. An example of that contrast is where liability was not imposed in a case where *D* carelessly spilt gasoline in a service station and a madman flicked a match into it precipitating a conflagration, but was imposed in a case where the defendant carelessly allowed prisoners to escape from custody and the prisoners damaged the yachts they stole to use in their escape.

A third difficulty with Hart and Honoré's linguistic analysis methodology is, ironically, that

there is virtually nothing in the language of the courts to support their contention that the law distinguishes cases such as the lightning cases on the basis of the law's notion of *causation*, let alone that the law explicitly imposed different *causal connection* requirements in the two cases. Hart and Honoré's linguistic 'analysis' was not based on a rigorous empirical survey, let alone such a survey across the full range of legal materials (Stapleton 2001*a*: 148–55): patterns of usage were merely asserted and illustrated. Yet counterexamples abound. For example, while Hart and Honoré assert that in cases of 'occasioning harm' the defendant has not 'caused' the harm (1985: p. xlv), Lord Reid pointedly noted in the yacht case:

It has never been the law that the intervention of human action always prevents the ultimate damage from being regarded as having been *caused* by the original carelessness ... every day there are many cases where, although one of the connecting links is deliberate human action, the law has no difficulty in holding that the defendant's conduct *caused* the plaintiff loss.[2]

A fourth and core problem with Hart and Honoré's approach is that it fails to segregate clearly contexts in which causal language was used by courts merely to communicate involvement under the causation element of a cause of action and contexts where a denial of proximate cause/scope was used to communicate a refusal to impose responsibility for the outcome with which the breach of obligation was clearly involved. Courts had long recognized these as two distinct enquiries (Williams 1961). Green appreciated the normative nature of the proximate cause/scope step but despaired of finding in case law the principles on which this truncation of legal responsibility was done. Hart and Honoré did see some patterns in case law concerning that step but were unwilling to declare that those patterns always tracked normative considerations. Thus even when Hart and Honoré *did* identify a factual feature, such as the intervention of lightning, often present when courts refuse to impose liability for want of proximate causation, they failed to explore the *normative* basis for these truncations of responsibility (on which see below), preferring merely to label them 'causal' and 'eminently suitable for submission to a properly instructed jury' (1959: 275).

Yet distilling normative principles from the proximate cause/scope case law can be done and, in the interests of legal clarity, should be done. For example, the damage by lightning in the ambulance case is a coincidental consequence of the speeding: in other words, speeding does not generally increase the probability of lightning injuries. The law may well judge it inappropriate (because it is unfair or because it does not directly deter the conduct of concern to the law, namely inadvertent speeding) to impose liability for the coincidental consequences of a breach of obligation. In contrast, damage due to the fumes ignited by lightning is not coincidental. This provides a coherent normative rationale for the divergent results in the cases and one on which a lawyer would be able to advise clients as to future conduct. Yet, regrettably, this rationale is masked by Hart and Honoré's preferred explanation: that in the ambulance case the tort of negligence requires, *for some unstated reason*, that the plaintiff prove that the breach satisfied the central 'causal connection' to the injury (which it does not because lightning intervened) but that, *again for some unstated reason*, the tort of negligence does not require the plaintiff in the fumes case to prove that central type of 'causal connection'. Predictably, this obscuring 'word magic' attracted strong criticism from US

Realists such as Leon Green (1962). Its dangers for clarity in legal reasoning are clear.

A fifth problem in Hart and Honoré's account of causation in the law is that it rests on a snapshot of causal usage frozen in the late 1950s. Their assertion that 'ordinary language'/'common sense' causal principles are 'facts' (1959: 86) does not adequately accommodate the late-Wittgensteinian insight that meaning cannot be divorced from the activities of the language users. Yet the language-games of lawyers are clearly embedded in a social practice that is in constant flux (Bix 2005: 223). Legislators (whose modern regulatory enactments Hart and Honoré by and large ignore) and courts change the pattern of legal obligations over time and such normative developments affect the sorts of conduct that may be prohibited or mandated. As we have seen, this breach issue in turn affects the involvement issue (because it changes the content of the hypothetical no-breach worlds to which the actual world is compared) and thereby causal usage. In short, the pattern of causal usage in the law, even if it does reflects ordinary language usage and as such is a 'social fact' at any one point of time (Lucy 2007), is contingent on the evolution of legal norms.

A good illustration is the greater willingness of modern courts and legislators to mandate affirmative action to prevent the intervention of a deliberate wrongdoer. Today the law might mandate that a store provide lighting in its surroundings to deter criminal attacks on its customers. When a customer, attacked in the unlit grounds of the defendant store in 2008, succeeds in his action in the tort of negligence, there will be a legal finding that the omission of lighting by the store was a cause of the customer's injury. In 1900 the law did not mandate such conduct so there would have been no such causal usage at that time.

Similarly, liability based on wrongfully providing reasons, long recognized in areas such as deceit, has also burgeoned with the recognition in 1963 that merely negligent advice resulting in economic loss may be actionable. As Hart and Honoré predicted, the more defendants are held liable in circumstances such as these and the affirmative duty contexts, the less it is 'helpful to describe the law with reference to the common-sense notions expounded' in their book, based as they were on legal usage of causal terminology in the third quarter of the twentieth century (1959: 180, see also 172).

4.3 Corrective Justice

Michael S. Moore (1999: 4) believes that the 'best goal for tort law' is corrective justice, which mandates that 'legal liability tracks moral responsibility'. But if in tort law '"[c]ause" has to mean what we mean when we assign moral responsibility for some harm, and what we mean in morality is to name a causal relation that is natural and not of the law's creation', (ibid., emphasis added), then corrective justice 'demands a robustly metaphysical interpretation' of 'cause'. In particular, corrective justice needs a metaphysical account of the sudden truncation of liability under the label of proximate cause/scope untainted by the influence of non-metaphysical considerations (consequentialist policy concerns).

Moore argued that the Hart and Honoré analysis of when we *name* relations as 'causal' was incomplete because it lacked such a firm pre-legal, 'plausible, understandable, communicatable, metaphysical' (1999: 2) basis for its notion of causation (2000: 854–5). Yet Moore adopts an odd strategy to find a pre-legal notion of cause: he constructs an account that

inexplicably draws on alleged patterns of causal usage (in the law no less!) which he seems to assume reflect moral concerns, and then claims this provides an avenue to a sound *metaphysical* account of causation. Here is that strategy.

Moore notes that the law's usage of causal terms, particularly in the area of proximate causation/scope, seems to presuppose that its concept of 'causation' must meet certain requirements. Moore (1999: 45) enquires 'whether there is any metaphysical theory of causation that can endow causation with' such requirements'. By taking at face-value the usages of 'cause' in liability doctrines, he deduces for example: that 'the law's concept of cause presupposes that causation both tapers off over time and breaks off suddenly at certain points in time' (ibid. 9) and that 'increased culpability has been treated as a kind of aphrodisiac to causation, enhancing the latter's reach and power' (ibid. 27).

Moore then argues that a metaphysical account of the concept of causation presupposed by the law can only be achieved if that concept is pruned. He claims that '[t]he law has mixed too many extraneous elements into what it calls 'causation' for there to be much hope for any metaphysical translation' (ibid. 28). For this reason Moore argues that we need to ignore certain doctrines on the basis: that 'they cannot be doctrines of cause-based liability, despite their self-labelling in these terms' (ibid. 6); and they make demands on the concept of causation that are 'obviously impossible ones for any metaphysics to meet' (ibid. 28). One example of the law making such a 'mistake' is when it seems to demand that 'causation be a relation affected by the degree of culpability with which the act (that is the putative cause) was done' (ibid.). Moore argues that 'there is no metaphysical account of causation that could meet this demand' and that '[c]ausation cannot be a real relationship in the world and [at the same time] be influenced by ... culpability' (ibid. 35). Another example is the treatment of the intervention by voluntary human conduct under the rubric of proximate cause/scope about which Moore concludes that 'it is hard to see how metaphysics can explain these legal discriminations' (2000: 877; Hurd and Moore 2002: 405).

A central flaw in Moore's approach is that he does not clearly explain, let alone justify, the criteria by which he chooses to prune 'the concept of causation presupposed by the law', and he does not confront the problem that, once he has effected such a selective pruning, it is not clear what the nature of his project has become. Moore (2000: 857) himself dismisses ordinary language philosophy because '[i]t allows the nature of the thing, causation, to be fixed by the conventions of present usage' and accepts that '[t]he nature of causation—"what causation is"—is a matter of fact, inviting theoretical speculation' (ibid. 855). He then presents what he claims is a 'plausible' metaphysical theory of causation having made two major but inexplicable normative moves. First, he now relies on intuition to make the normative choice of not allowing an omission to qualify as a cause: since we draw a moral 'distinction between our responsibility for making the world worse and our responsibility for making it better ... [t]he easiest, most *intuitive* way to draw this distinction is by using causation to mark the difference' (1999: 31–2, emphasis added). Moreover, he asserts that, where there is liability for negligent provision of an opportunity for another to do harm, 'the liability is not *cause*-based liability ... [because] these are cases of true omission liability' (ibid. 36, emphasis added).

Secondly, a central claim within the metaphysical theory of causation that Moore provides is that causal relations peter out gradually by transmission through events. 'This is because

causation is a scalar relation (a more-or-less affair) and because the degree of causal contribution by some act to some harm becomes less and less as successively larger groups of other events join the act in causing the harm' (Hurd and Moore 2002: 410). On this view he claims that at least one limit on liability imposed under the label of 'proximate cause' ('scope'), spatio-temporal proximity, can be rationalized as a 'good proxy for this progressive diminishment in causal contribution' (ibid.). But Moore (2000: 828) had deduced the scalar nature of his 'prelegal' metaphysics of causation from patterns of causal usage in legal materials: why was not it pruned along with the 'erroneous' idea held by courts and legislatures that omissions can have causal status?

The incoherence of Moore's approach results from his apparent embrace of a physicalist approach (ibid. 877) that sees causation as 'a real relationship in the world'; in combination with a respect, albeit selective, for linguistic usage, specifically that with a scalar aspect, plus an associated failure to acknowledge and deconstruct the atypical nature of the usage of causal terms in the criminal law. As Stephen J. Morse (2000: 880) notes, it is metaphysically implausible that, within 'the universe's ontology of physical cause and effect ... there are "sharp breaks" in the "causal chains" of the universe that would provide a moral rationale for the same sharp breaks in [proximate cause] doctrine that Michael accurately identifies'. Rather, '[t]he best understanding ... is that causation is a seamless web ... [T]here are no gaps or sharp breaks in causation ... [C]ausation just keeps rolling along' (ibid. 889). In any case, the dominant view of modern lawyers is that the 'proximate *cause*' terminology of the past was highly misleading and masked complex normative judgements about the appropriate scope of legal responsibility for consequences of conduct, and that these judgements (unlike the metaphysical reality of the world) change over time and between jurisdictions for a variety of reasons. To the extent that something seems to be cut off or peter out under this legal label it is legal responsibility.

4.4 Lawyer-Economists: Marginalized Causation

In the last four decades a particularly influential account of the law has been that of the legal economists who have reconceptualized the law from the perspective of the efficient allocation of resources and maximizing or minimizing certain behaviours. The building blocks of the approach are thus: assume the law seeks to support the most efficient allocation of resources, namely that entitlements are in the hands of those who value them the most (judged by capacity and willingness to pay). In an environment of zero transaction costs, that state of optimal efficiency will result whatever the distribution of initial legal entitlements, because interested parties will bargain around those entitlements: this is the Coase Theorem (Coase 1960). Where, however, there are significant transaction costs, as is usually the case, the initial legal allocation of rights becomes crucial to securing this state of optimal efficiency; so the lawyer-economist's normative agenda is to allocate initial legal rights to those who value them the most.

For example, the interaction of hunters and walkers involves a risk of personal injuries to the walker. Hunters and walkers both value their activity. Were they able to transact we might find that the most efficient state of affairs is for both to continue but for the hunters to take

care where they shoot (because the walkers would be willing to pay a large enough 'bribe' for the hunters to do so). The reality of transaction costs that prevent this bargaining then justifies a liability law that gives the mountain walker an entitlement not to be injured by the carelessness of the hunter because this would give hunters the appropriate, 'efficient' incentive to take care. Conversely we might find that the efficient state of affairs is one where hunters need take no care of walkers, in which case the lawyer-economist would not support liability on the hunter, thus leaving the walker with an incentive to avoid the risk associated with his interaction with the hunter. In short, for the lawyer-economist liability is a mere instrumental device to generate incentives for future conduct.

In the economic account, past injuries are 'sunk costs' and how they came to occur is of no direct interest. In any case, when a hunting accident occurs the activity of the hunter and the activity of the walker are both involved factors so, for the lawyer-economist, 'causality is reciprocal'; and it is a crude nonsense to talk about internalizing social costs to the activity that 'causes' them (Coase 1960: 28–42; Calabresi 1961). Thus Landes and Posner (1983: 131) assert that 'causation in the law is an inarticulate groping for economically sound solutions' and argue:

> If the basic purpose of tort law is to promote economic efficiency, a defendant's conduct will be deemed the cause of an injury when making him liable for the consequences of the injury would promote an efficient allocation of resources to safety and care; and when it would not promote efficiency for the defendant to behave differently, then the cause of the accident will be ascribed to an 'act of God' or some other force on which liability cannot rest. (ibid. 110)

> In Calabresi's (1975) terminology, only if the defendant would be the 'cheapest cost avoider' of such injuries in the future, should he be identified by the law as a 'cause' of the past injury.

The fact that efficiency theories of law cannot adequately explain or justify the causation requirement in legal doctrine is widely regarded as a serious flaw in the approach. Moreover, the view that 'causation in the law' is and should be a mere instrumental label leads Calabresi to a major departure in causal theory. Where a speeding motorist breaks the leg of a pedestrian who is then killed when his ambulance is struck by lightning, the death is a coincidental consequence of the speeding: speeding does not generally increase the probability of lightning injuries. By definition, where an activity does not generally increase the probability of an outcome, there is no way in which an actor can alter the way he engages in that activity to prevent the outcome. It is, therefore understandable why, on efficiency grounds, Calabresi argues against liability in such circumstances. To prevent liability being imposed in such circumstances Calabresi (ibid. 71) invents a 'causal' requirement of 'causal linkage': '[t]here is a causal link between an act or activity and an injury when we conclude on the basis of the available evidence that the recurrence of that act or activity will increase the chances that the injury will also occur.' Since speeding does not increase the chance of lightning strikes, there is no causal link between the speeding of the motorist and the pedestrian's death.

In terms of this chapter, what Calabresi in effect asserts is that the investigation underlying

causal usage in the law should be an interrogation that identifies as a cause only a factor that is involved in the existence of the particular phenomenon of interest *and* only where that phenomenon is a non-coincidental consequence of that factor.

Couching a no-liability-for-coincidence rule in terms of there being a requirement of causal linkage is, however, triply inconvenient for the law. It obscures Calabresi's normative reason for no-liability in the ambulance case: that liability for coincidental consequences of careless conduct such as inadvertent speeding cannot directly deter the conduct of concern to the law, namely speeding. Moreover, it would be as problematic as the approach of Hart and Honoré in cases where the law *does* seek to impose liability for coincidental consequences. For example, in a case of wrongdoing that requires intention, such as deceit, there is a sound reason *for* imposition of liability for coincidences. This is an efficiency reason with which Calabresi should be highly sympathetic, namely that such liability can have second-order deterrent effects by encouraging these necessarily advertent actors to review downwards their future activity level in engaging in such wrongful conduct. Finally, since it suggests that the focus of an efficiency analysis should be on 'ex ante probabilistic linkage (increased risk) analysis' (Wright 1985*b*: 453; see also Wright 1987) it fails to account for the fact that liability is typically limited to contexts in which some form of injury has been suffered.

4.5 The New Realism

Though Hart and Honoré's linguistic approach proved inconvenient for lawyers, they provided a major advance in 1959 by formulating an algorithm to identify which factors might qualify as a 'causally relevant condition'. From the mid-1980s Richard Wright developed and popularized this NESS algorithm as a test for causation in the law, preferring its formulation to the INUS algorithm suggested in 1965 by Mackie (Wright 1988: 1023 n. 113). Wright asserts, '[the] basic concept of causation, which we all intuitively employ, is formalized in the NESS test, which in its full form states that a condition contributed to some consequence if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of the consequence' (2003: 1441).

We have already noted an illustration of the NESS test: in the hunters' case a subset of factors comprised of all antecedent factors except the carelessness of Hunter No. 2 was *sufficient* for the occurrence of the death of the walker and the carelessness of Hunter No. 1 was *necessary* for the sufficiency of that subset (i.e. in the hypothetical world consisting of this subset minus the breach by Hunter No. 1 the death would not have happened); his carelessness was a NESS factor.

It is generally accepted by interested academic lawyers in the United States that using NESS (necessary element for the sufficiency of a sufficient set) as a test for causation seems, as a practical matter, to yield a coverage that is sufficiently comprehensive and 'factual' to satisfy the wide projects of the law. For example, the NESS test can accommodate mere conditions, omissions, and the communication of information as causes. It also deals smoothly with, and identifies as causes duplicate factors that are involved in an overdetermined outcome such as the death of the mountain walker, as we have just seen. It allows the fact that the analytically prior normative stage of breach determines which hypothetical no-breach worlds are of legal

interest and it is a test not tainted by ideas of breaks in causation or causation petering out.

Nonetheless, critics have claimed that there are residual problems with Wright's otherwise impressive exposition of the NESS device. Some criticisms are relatively minor such as concern with Wright's occasional appeal to intuitions (e.g. Wright 1988: 1003, 1009) and his curious assertion that, in as much as NESS requires a 'counterfactual' analysis, this is to proceed not by speculating on what might have happened in a hypothetical world where, *inter alia*, the specified factor is absent, but by considering what actually happened (Wright 1985*a*: 1806–7; 2003: 1445 n. 67; 2007: 296).

But one criticism of Wright, that by Fumerton and Kress, is fundamental and devastating. They demonstrate that, far from delivering on his claim that NESS 'is not just a test for causation, but is itself the meaning of causation' (Wright 1985*a*: 1802), Wright's reliance in NESS on the idea of 'causal sufficiency' (Wright 2001: 1103 n. 113) involves 'vicious conceptual circularity' (Fumerton and Kress 2001: 84). How can NESS capture the 'meaning' of causation, if it is dependent on external 'causal laws'?

This attack is indeed fatal to Wright's claim concerning the *character* of the NESS test. Nevertheless, it does not, in my view, detract from the potential *practical* value of NESS to the lawyer. Let me explain by recapping the approach I advocate.

I have argued that when we investigate the world we must choose between a variety of possible limited interrogations: blame, explanation, physical role, involvement, and so on. Since the different results of these interrogations have often been expressed in causal terms, there can be no coherent account of causation in the law until we have *first* chosen which interrogation *should* be the one to underlie causal terms in the law.

Next I argued that this interrogation should be chosen to meet the wide needs of the law. These range from the conceptualization of a farmer's blameless omission in breach of contract 'causing' a loss to a retailer, to determining the most efficient regulation of the risks of hunting accidents, and to identifying the role played by a single vote within a unanimous vote to expel a club member.

In all these projects the law must be able to identify whenever a specified factor was involved in the existence of a particular phenomenon of interest, where the notion of 'involvement' identifies that there is a contrast between the actual world and some hypothetical world from which we exclude (at least) that specified factor: this contrast being that, while in the former world the phenomenon exists, in the latter it does not. We can generate such contrasts of necessity in three ways. For example, when there is this contrast between the actual world and a hypothetical world from which we simply exclude the specified factor, we can convey this information by saying that the factor was involved in the existence of the phenomenon by being necessary for it. When there is this contrast between the actual world and a hypothetical world from which we exclude both the specified factor and a duplicate factor, we can convey this information by saying that the factor was involved in the existence of the phenomenon by a relation of duplicate necessity. In a similar way we can identify a third form of involvement, namely mere 'contribution' to the existence of the phenomenon.

Next I have argued that the most convenient choice of interrogation to underlie causal usage in law is this one of involvement (because it most conveniently enhances clarity of legal analysis): whether a specified factor was involved in the existence of a particular phenomenon

in any of these three ways. Now we can see the true value of the NESS idea: not as a self-evident 'meaning of causation', but simply as an extremely effective algorithm for identifying all the relationships of involvement (between a specified factor and the existence of a particular phenomenon) with which the law must deal and which I have argued should be *chosen* as the meaning of causation in the law.

Because this determination of involvement will be made using our knowledge of the physical laws of nature, evidence of behaviour, and so on, choosing 'involvement' as the meaning of 'causation' in law carries the potential for the concept of causation to be untainted by normative controversies. This would mean, for example, that where all the facts of the case are known, there would be no room for disagreement on the issue of causation.

It is Wright's failure to accept that NESS is merely a device we can manipulate to serve the role we have previously decided to assign to it that has led to incoherence in his account, specifically in relation to the issue of disaggregation and in relation to how NESS is to be applied to cases of dependent double omissions (Stapleton 2008).

5. FUTURE DEBATES

While past controversy about the meaning of causation in the law is being resolved in favour of embracing the wide notion of involvement, for which NESS seems to provide a relatively satisfactory algorithm, far greater debate now centres around two other related issues. First is the problem of where and why the law should relax its rules of proof in relation to this notion of involvement to assist, for example, a plaintiff who would otherwise face insuperable evidentiary gaps.

The second area of intense legal interest is the appropriate scope of responsibility for the consequences of wrongful conduct. The normative concerns governing this scope issue have not yet been adequately investigated, hindering clarity in judicial reasoning and therefore producing uncertainty in the law.

FURTHER READING

Hart and Honoré (1985) is the most striking work in the field. Simply by virtue of being the most comprehensive linguistic analysis in any field, it deserves attention. Nonetheless its philosophical complexity has prevented it having wide influence in legal circles. A more efficient introduction to modern issues concerning causation in the law is Wright (1988), which deals with most earlier contributors and provides many illustrations from case law. The philosophical naivety of key claims made by Wright for the NESS test are exposed in Fumerton and Kress (2001), Fischer (2005–6), and Stapleton (2008) (which contains the first significant exposition of the 'involvement' approach to causation in the law).

REFERENCES

American Law Institute (2005). *Restatement of the Law (Third) of Torts: Liability for*

Physical Harm (Proposed Final Draft, April 6, 2005). Philadelphia: American Law Institute.

BECHT, A. C., and MILLER, F. W. (1961). *The Test of Factual Causation in Negligence and Strict Liability Cases*. St Louis: Washington University.

BIX, B. (2005). 'Cautions and Caveats for the Application of Wittgenstein to Legal Theory', in J. K. Campbell, M. O'Rourke, and D. Shier (eds.), *Law and Social Justice*. Cambridge, Mass.: MIT.

BRUDNER, A. (1998). 'Owning Outcomes: On Intervening Causes, Thin Skulls, and Fault-Undifferentiated Crimes', *Canadian Journal Law and Jurisprudence* 11.

CALABRESI, G. (1961). 'Some Thoughts on Risk Distribution in Torts', *Yale Law Journal* 70: 499–553.

—— (1975). 'Concerning Cause and the Law of Torts: An Essay for Harry Kalven, Jr.', *University of Chicago Law Review* 43: 69–108.

COASE, J. (1960). 'The Problem of Social Cost', *Journal of Law and Economics* 3: 1–44.

Commonwealth of Australia (2002). *Review of the Law of Negligence: Final Report* ('Ipp Report'). Canberra: Commonwealth of Australia.

DOBBS, D. (2000). *The Law of Torts*. St Paul, Minn.: West Group.

FISCHER, D. A. (2005–6). 'Insufficient Causes', *Kentucky Law Journal* 94: 277–317.

FOOT, P. (1963). 'Hart and Honoré: Causation in the Law', *Philosophical Review* 72: 505–15.

FUMERTON, R., and KRESS, K. (2001). 'Causation and the Law: Preemption, Lawful Sufficiency, and Causal Sufficiency', *Law and Contemporary Problems* 64: 83–105.

GREEN, L. (1927). *The Rationale of Proximate Cause*. Kansas City: Vernon Law Book Co.

—— (1928–9). 'Are There Dependable Rules of Causation?', *University of Pennsylvania Law Review* 77: 601–28.

—— (1962). 'The Causal Relation Issue in Negligence Law', *Michigan Law Review* 60: 543–76.

—— (1972). 'Identification of Issues in Negligence Cases', *Southwestern Law Journal* 26: 811–29.

HALL, E. J. (2004). 'Two Concepts of Causation', in J. Collins, E. J. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, Mass.: MIT, 225–76.

HANCOCK, R. (1961). 'Books Reviewed', *Natural Law Forum* 6: 143–52.

HART, H. L. A., and HONORÉ, A. M. (1959). *Causation in the Law*. Oxford: Oxford University Press.

—— (1985). *Causation in the Law*. 2nd edn. Oxford: Oxford University Press.

HURD, H. M., and MOORE, M. S. (2002). 'Negligence in the Air', *Theoretical Inquiries in Law* 3: 333–411.

LANDES, W. M., and POSNER, R. A. (1983). 'Causation in Tort Law: An Economic Approach', *Journal of Legal Studies* 12: 109–34.

LEITER, B. (1999). 'Positivism, Formalism, Realism', *Columbia Law Review* 99: 1138–64.

—— (2005). 'American Legal Realism', in W. Edmundson and M. Golding (eds.), *The Blackwell Guide to the Philosophy of Law and Legal Theory*. Oxford: Blackwell.

LIPTON, P. (1992). 'Causation Outside the Law', in H. Gross and R. Harrison (eds.), *Jurisprudence: Cambridge Essays*. Oxford: Oxford University Press, 127–48.

LUCY, W. (2007). *Philosophy and Private Law*. Oxford: Oxford University Press.

MACKIE, J. L. (1965). 'Causes and Conditions', *American Philosophical Quarterly* 2: 245–64.

—— (1974). *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.

MOORE, M. S. (1999). 'Causation and Responsibility', *Social Philosophy & Policy* 16: 1–51.

—— (2000). 'The Metaphysics of Causal Intervention', *California Law Review* 88: 827–77.

MORSE, S. J. (2000). 'The Moral Metaphysics of Causation and Results', *California Law Review* 88: 879–94.

OLIPHANT, H. (1928). 'A Return to Stare Decisis', 14 *American Bar Association Journal* 71–6.

ROBERTSON, D. W. (1978). 'The Legal Philosophy of Leon Green', *Texas Law Review* 56: 393–437

—— (1996–7*a*). 'The Common Sense of Cause in Fact', *Texas Law Review* 75: 1765–800.

—— (1996–7*b*). 'Allocating Authority among Institutional Decision Makers in Louisiana State-Court Negligence and Strict Liability Cases', *Louisiana Law Review* 57: 1079–117.

—— (1997). 'The Vocabulary of Negligence Law: Continuing Causation Confusion', *Louisiana Law Review* 58: 1–33.

SCHAFFER, J. (2005). 'Contrastive Causation', *Philosophical Review* 114: 327–58.

STAPLETON, J. (2001*a*). 'Unpacking Causation', in P. Cane and J. Gardner (eds.), *Relating to Responsibility*. Oxford: Hart, 14–85.

—— (2001*b*). 'Legal Cause: Cause-in-Fact and the Scope of Liability for Consequences', *Vanderbilt Law Review* 54: 941–1009.

—— (2003). 'Cause-in-Fact and the Scope of Liability for Consequences', *Law Quarterly Review* 119: 388–425.

—— (2008). 'Choosing What We Mean by "Causation" in the Law', 73 *Missouri Law Review* 433–80.

TADROS, V. (2005). *Criminal Responsibility*. Oxford: Oxford University Press.

WILLIAMS, G. (1961). 'Causation in the Law', *Cambridge Law Journal*, 62–85.

WRIGHT, R. W. (1985*a*). 'Causation in Tort Law', *California Law Review* 73: 1735–828.

—— (1985*b*). 'Actual Causation vs. Probabilistic Linkage: The Bane of Economic Analysis', *Journal of Legal Studies* 14: 435–56.

—— (1987). 'The Efficiency Theory of Causation and Responsibility: Unscientific Formalism and False Semantics', *Chicago-Kent Law Review* 63: 553–78.

—— (1988). 'Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts', *Iowa Law Review* 73: 1001–77.

—— (2001). 'Once More into the Bramble Bush: Duty, Causal Contribution and the Extent of Legal Responsibility', *Vanderbilt Law Review* 54: 1071–132.

—— (2003). 'The Grounds and Extent of Legal Responsibility', *San Diego Law Review* 41: 1425–531.

—— (2007). 'Acts and Omissions as Positive and Negative Causes', in Neyers, J., Chamberlain, E., and Pitel, S., *Emerging Issues in Tort Law*. Oxford: Hart, 287–307.

INDEX

Note: page numbers in italic refer to figures.

[1] Thanks to Ursula Coope for helpful comments on a previous draft. These are the masculine and neuter forms.

[2] *Phaedo* 100b-101c. Alternative translations are 'partakes of', 'shares in'.

[3] It was natural, juridically, to say that *X* (usually a person) *has* the *aitia* for *Y;* explanatorily, that *X is* the *aitia* for *Y.*

[4] *episteme* is sometimes translated 'understanding'. Aristotle's theory of *apodeixis* has sometimes been misunderstood and castigated as a theory of scientific discovery, according to which we start with the premisses, which are truths already known, and deduce syllogistically that *p,* where *p* is supposedly new information. But in explanatory *apodeixis* the movement is in the opposite direction. Beginning from a *p* that first figures as an accepted truth requiring to be explained, we seek suitable premisses from which to deduce it.

[5] The relevant part of premiss (1) for A is: <u>what does not twinkle is near</u>.

[6] For Aristotle, necessary truth is not limited to what can be known a priori.

[7] *Nicomachean Ethics* 6.7, 1141a19-20, with 6.3 and 6.

[8] The formula of the essence of *X* is the fundamental structure of *X* on which its other essential properties are grounded. The essence of a musical interval is the ratio between the lengths (e.g. of a string) required to produce the notes. The musician imposes the ratio on a 'matter' consisting of (e.g.) his string-length, just as the sculptor imposes the structure of the statue on the bronze.

[9] The four types of cause illustrated here came to be known as, respectively, the *material, formal, efficient,* and *final* causes.

[10] Aristotle, *Physics* 2.8–9.

[11] Evolution by natural selection is excluded from the account, because Aristotle assumes eternity of the species.

[12] *Metaphysics* 12.7.

[13] To understand this we must bear in mind that Aristotle's inanimate elements are homogeneous masses, structureless through and through. Any theory that inanimate materials can *of themselves* combine into complicated biological formations is almost certain to try to make sense of this by postulating complex powers arising from microstructure.

[14] 'Necessary connexion' is one of the three elements into which Hume analyses the causal relation.

[15] In *Metaphysics* 2.2 Aristotle argues against an endless series under each of the four headings. For chains of causal dependence between contemporaneous items, see *Physics* 7.2 and 8.5, 256a4-256b27.

[16] Nothing here depends on any difference of meaning between 'freedom' and 'responsibility'. Moreover, 'behaviour' and 'action' here include choice, decision, and willing.

[17] A more sophisticated version of this sort of move, involving (as reported by Cicero) a distinction between 'complete and primary' and 'auxiliary and proximate' causes, was proposed by the Stoic Chrysippus in a bid to reconcile freedom and responsibility with the Stoic doctrine of 'fate' (= universal determinism). I am only not responsible if the complete and primary cause of my behaviour is other than, or outside, me. See Long and Sedley (1987), §§55.I and 62.C, with commentary, pp. 392-3.

[18] Cf. R. Chisholm, 'Human Freedom and the Self' (Lindley Lecture 1964, University of

Kansas) §11: 'some would attribute only to God [the status of] a prime mover unmoved'. Chisholm extends the status to free human agents; he does not discuss non-human agent-causes. His account, unlike the present, assumes that the existence of unmoved movers is incompatible with determinism.

[19] *Physics* 7.1; 8.4-5.

[20] In *Physics* 8.4-5 Aristotle recognizes a category of things that move themselves, but he then argues that these are wholes made up of distinct factors, mover and moved (ibid. 8.5, 257a31 ff.)

[21] Ibid. 7.1; 8.5.

[22] Cf. *Metaphysics* 7.7. Although the form in the craftsman's soul is what we might call 'mental' (since possessing this sort of form is an intellectual attainment), Aristotle invokes it partly to illustrate the case of biological substances, whose inner vitalistic principle or soul contains, or even is, a *non-mental* active form.

[23] A skill in the sense of a potential for skilled action becomes a skill-in-action (i.e. becomes activated) only if the agent desires to use it, and to use it in one way rather than another (ibid. 9.5).

[1] Hume 2000*a*/1739–40, henceforth cited as 'THN' and followed by book, part, section, and paragraph number. References to particular paragraphs are also followed by references to page numbers of Hume 1978*a*/1739–40 (cited as 'SBN').

[2] Hume 2000*b*/1748, henceforth cited as 'EHU' and followed by section and paragraph number. References to particular paragraphs are also followed by references to page numbers of Hume 1975/1748 and 1751 (cited as 'SBN').

[3] On the other hand, it would be *consistent* with Hume's theory if some members of the revival set of the concept of CAUSATION were themselves pairs of other *abstract* ideas—such as those of SMOKING and LUNG CANCER—in experienced users of the concept.

[4] Hume 1998/1751, henceforth cited as 'EPM' followed by section and paragraph number. References to particular paragraphs are followed by references to page numbers of Hume 1975/1748 and 1751 (cited as 'SBN').

[5] This is just one example. He also writes of 'proving' and 'knowing' the causal 'dependence' of ideas on preceding impressions (THN 1. 1. 1. 8; SBN 4–5), of 'the true and real cause of [an] idea, and of the belief which attends it' (THN 1. 3. 8. 8; SBN 102); of the sufficiency of constant conjunction to render something a 'real cause' (THN 1. 3. 14. 32; SBN 171); of having established 'the truth of my hypothesis' about the causes of causal inference itself (THN 1. 4. 1. 8; SBN 183–4); and of having 'proven' the causes of pride and humility (THN 2. 1. 12. 1; SBN 324–5).

[1] While it is true that Kant denies that we can prove God's existence by theoretical means in the Critical period, he does think that God's existence can be established on practical grounds, such that the main contention of the Principle of Coexistence can still be maintained (even if in a somewhat different form).

[2] This is not to say that Kant does not provide clarification of the precise meaning of the category of causality in other contexts.

[3] While one might attempt to shore up the argument, on straight philosophical grounds, with a verificationist premiss (such that the conceptual relationship between succession and causality can be meaningful only if they can in fact be found in experience), it seems dubious that the sceptic will see the need to accept such a controversial assumption. On textual grounds, one might appeal to further argument in the Refutation of Idealism, though the Second Analogy ought to stand or fall on its own, without supplementation by argument later in the text.

[4] There are further issues pertaining to mutual interaction that require clarification. For example, one might naturally wonder whether the activity of each substance by which it determines the state of the other must occur at the same time as the other substance's activity. If this were the case, it would then seem that one would have to know simultaneity in order to know the existence of mutual interaction, but since mutual interaction was supposed to be revealed as a necessary condition of simultaneity, the argument can seem to be circular.

[5] One can also be sceptical about the existence of natures, but if natures were to be admitted, then they can also play a significant role in the generation of necessary connections and in an explanation of the kind of necessity causal relations have, namely, natural necessity.

[6] These issues are discussed by Watkins (2005).

[7] One can thus see that there is a genuine basis in Kant's texts for Strawsonian style transcendental arguments (which attempt to explain the possibility of different kinds of experience such as of particulars).

[1] Where page numbers take the form x/y, 'x' refers to the page number in the original German, and 'y' to the page number in the English translation listed in the bibliography. Longer italicized passages of the originals are reproduced in roman unless the emphasis is relevant for the argument as presented here.

[2] The German *Gesetzmäßigkeit* is notoriously difficult to translate. Sometimes it is used as an umbrella term for causal law (*Gesetz*) and statistical regularity (*Regelmäßigkeit*), while on other occasions it—deliberately or uneasily—sits between them or denotes a mixture still to be sorted out. For this reason I translate on a case-to-case basis and add the German original.

[3] Cf. Schlick's correspondence in the Wiener Kreis Stichting, Rijksarchief Noord-Holland.

[1] Stalnaker (1968) and Lewis (1973*a*) develop the accepted semantics for counterfactuals in terms of similarity of possible worlds to the actual world. See the introduction of Collins, Hall, and Paul (2004) for a more developed discussion of the evaluation of counterfactuals in the context of a counterfactual theory of causation.

[2] Chalmers and Jackson (2001) hold that the success of an ontological reduction depends on the possibility of a certain sort of conceptual analysis.

[3] This idea is suggested by Karen Bennett in her forthcoming 'Two Causal-isms' and seems to be at work in discussions of trumping by Schaffer (2000*b*) and Lewis (2004*a*).

[4] Do not confuse the pragmatic approach with contextualist approaches (although the latter also face the objection that causation is conflated with causal explanation). Contextualist approaches are accounts where causal relations are indexed to contexts, much like velocities are indexed to frames of reference, and do not necessarily involve subjective or normative elements.

[5] I discuss some of these issues in more detail in 'The Handmaiden's Tale' (forthcoming).

[6] Lepore and Loewer (1987) argue, relatedly, that since dependence is sufficient for causation we can infer the presence of mental causation. See also Bennett (2003).

[7] Lewis (1973*b*) argues that a sufficiency account of causation, where C is a cause of E iff C is sufficient, under the laws, for E, cannot handle cases of pre-emption. Hall and Paul (forthcoming) argue that we should be cautious about the claim that the counterfactual analysis is an obvious winner with respect to the treatment of these sorts of cases.

[8] The sufficiency theorist may be able make use of derived laws, but must then explain the difference between derived laws and unlawful generalizations.

[9] Schaffer (2000*a*) describes how the interior mechanism of the firing of a gun, which might seem like a paradigm case of an uninterrupted causal process, actually involves a causal chain that includes double prevention.

[10] Terminological note: Lewis (1986*a*; 2000; 2004*a*) uses 'late preemption' differently from how I am using it: for Lewis, it refers only to cases where the causal chain is interrupted because the occurrence of the effect (caused by pre-empting cause *C*) prevents events in the back-up chain from occurring.

[11] Note that interpreting Figure 8.5 in this way involves a slight change in how the neuron diagram is being used—it represents the causal story at a time rather than over time.

[12] Indeterministic laws might not be violated, since it is consistent with such laws to hold that extremely improbable events are possible. So one might hold that it is extremely improbable yet physically possible that, in a case of pure overdetermination, an effect (such as the shattering) would occur just as it would have had it been caused by a single rock throw. However, the problem with additivity remains. In the case of indeterministic laws, the question we should ask is why the presence of an additional, overdetermining cause would not change the probabilities of the outcome. How could one lawfully claim that the probability of the shattering occurring (just as it actually did) is exactly the same whether there are two rocks or one? (Thanks are due to Jenann Ismael for raising the possibility of indeterminism in conversation.)

[13] The thesis about the dependence of the character of a causal chain on intrinsic features plus the laws is defended in Hall (2004).

[14] Even this may not be enough: in the end, in order to avoid esoteric counterexamples, the counterfactual analyst may be forced to rely on definitions involving intrinsicness, where, for example, the process between A and E and the process between C and E are required to exhibit certain intrinsic features or to intrinsically match certain causal processes. See Hall and Paul (forthcoming) for discussion.

[1] Reichenbach (1956) also develops a mechanistic understanding of causality. Cf. sect. 10.

[2] See e.g. Gillies (2000) for a survey of interpretations of probability.

[3] I am grateful to Donald Gillies for this suggestion.

[4] Good (1961*a*: A20, p. 313) appears to allow the possibility that common causes do not screen off their effects, but the account is fully developed only in the case in which they do (Good 1961*b*: }}9–11).

[5] I do not claim here that mechanisms are useful only for explanation and not for inference. Mechanisms also tell us about the stability of causal relationships, which is useful for inference. However, while difference-making is required for successful inferences, stability is not, since fragile causal relationships (i.e. those that are very sensitive to context) are useful for inferences in those contexts where they obtain. Similarly, difference-making has relevance for explanation as well as inference: in order to explain why an effect occurred it may be sufficient to point to those mechanisms that made a difference to the effect occurring. However, difference-making alone is not sufficient for explanation.

[6] This ideal set of evidence is not language-relative: given this evidence, there should be nothing left to know about physical reality, so that if these facts were expressible in some language, this language would have to be ideal inasmuch as it would need to be able to express all facts about physical reality.

[7] I am grateful to Bert Leuridan for discussion of this objection.

[8] See Williamson (2006*b*) for other desirable characteristics of a philosophical theory of causality.

[1] Some of Menzies and Price's language suggests such a view and it is also advanced by other writers. For example Hausman (1998) suggests that this is the origin of our idea of causal necessity.

[2] That is, on the projection hypothesis there is an obvious sense in which all causal claims involving inanimate objects are in error. We thus need an account of why it is justifiable to think of some of these claims as 'true' while others are false.

[3] There is a great deal of evidence that specialized neural systems are involved in the sense of agency or ownership of action and in the attribution of mental states to others and that these are largely distinct from the systems involved in understanding the behaviour of inanimate objects. On one natural construal of the projection hypothesis, it predicts, on the contrary, that the same areas are active both in theory of mind/attribution of agency and in the attribution of causal influence to inanimate objects. My reading of the available evidence is that there is relatively little overlap. Most studies show that attribution of agency/mental states involves the insula, anterior cingulate cortex, superior temporal sulcus (STS), temporal poles, ventromedial prefrontal cortex, perhaps amygdala to some extent. Causal attribution to inanimate objects involves V5/MT, STS, and some parietal areas in the case of Michottean, launching-style causal interactions, and dorsolateral PFC in the case of more abstract causal learning. The projection hypothesis also seems to predict that subjects (e.g. high-functioning autistics) who have deficits in mental state attribution would also have difficulties with causal attribution involving inanimate objects. The available evidence seems to contradict this prediction.

[4] A broadly similar framework is developed in Spirtes, Glymour, and Scheines ([1993] 2000). My understanding is that these latter authors were the first to introduce the 'arrow breaking' or 'equation wipe out' conception of interventions into current discussion.

[5] In addition to the considerations described below, the idea (and indeed the whole notion of an intervention) is broadly suggestive of Lewis's (1973) account of causation in terms of counterfactuals, the antecedents of which are made true by miracles. Readers interested in a more systematic exploration of these similarities (and some differences) are referred to Woodward (2003).

[6] As Woodward and Hitchcock (2003) observe one natural way of representing this is to think of the intervention variable as a 'switch' variable. For some range of values of the intervention variable I (values for which I takes an 'off' value) the variable intervened on, X, is a function of its parents and the value of I. For other values of I (values for which I is 'on'), the value of X is a function of the value of I alone.

[7] For a discussion of such connections, see Woodward (2003: ch. 7). One natural formulation is: If $P(Y/do\ X) = P(Y)$, then $P(X/parents(X)) = P(X/Parents\ (X).Y)$. This is one way of stating the so-called Causal Markov Condition.

[8] Some additional reasons are described in Woodward (2003: ch. 3).

[9] Spirtes, Glymour, and Scheines (1993) call this a failure of 'faithfulness'.

[10] This distinction is drawn in Pearl (2000)and Hitchcock (2001*b*), as well as Woodward (2003).

[11] Proposals along these lines are given in Halpern and Pearl (2001), Hitchcock (2001*a*), Woodward (2003). For an improved proposal that addresses some shortcomings in these

previous accounts, see Hitchcock (2007*b*).

[12] Another related argument concerning reduction is raised by the common claim that counterfactuals (including interventionist counterfactuals) as well as causal claims cannot be 'barely true' but instead require non-modal truthmakers of some kind—laws of nature being standard candidates for this role. If so, the argument continues, we should we appeal directly to this non-modal notion of law rather than to interventionist counterfactuals to explicate causal claims. Space precludes detailed discussion but note that this argument assumes that it is possible to explicate the notion of a law without appeal to interventionist counterfactuals and then use the former to ground the latter. Woodward (2003) denies this, arguing that invariance is the key feature of laws, where invariance is a counterfactual notion that has to do with stability under possible changes and is not reducible to non-modal notions.

[13] Norton (2007) provides one statement of the dominant view. In contrast, Frisch (2002) argues for an interpretation of classical electrodynamics that relies on 'rich causal assumptions', understood in explicitly interventionist terms.

[14] Yet another possible position would be to hold that causal claims play a central role in fundamental physics but that for the reasons described above, interventionist accounts fail to capture this role. However interventionist accounts are successful at elucidating causal claims in the special sciences. On this view, causal claims in fundamental physics would need to be given some other, non-interventionist elucidation.

[15] For more detailed discussion, see Woodward (2007*b*).

[1] Scriven (1966) invokes pre-emption cases in his defence of anti-reductionism. About an appeal to intermediate links, he says, 'This test does not apply where no such links are known, and since it is not logically necessary that there be any ... the test is not part of the meaning, of course' (ibid. 259–60).

[2] Woodward tentatively recommends taking this attitude towards trumping pre-emption. He does not even tentatively recommend it be taken towards any of his underdetermination cases. Nevertheless, the quotation voices well the kind of attitude many philosophers have taken towards the underdetermination examples.

[3] Is this too quick? Doesn't the fact that Vesuvius caused the destruction commit us to Vesuvius, the destruction, *and the causal relation between them*? No, it doesn't. Predications do not commit us to the existence of any properties or relations so long as we resist the urge to give some fully general analysis of predication. (The same goes for our use of sentential connectives such as 'because', 'before', 'and', ...) For the judicious metaphysician, it is the singular terms and the quantifications alone that can threaten to overpopulate an ontology. See Quine (1948); also see Devitt (1980).

[1] It is standard for causal models to leave the values of exogenous variables unspecified, and including equations only for the endogenous variables. This certainly makes sense if one intends the model to apply in a variety of contexts where the causal structure is the same, but the values of the exogenous variables differ. In this framework, the effects of interventions and the truth values of interventions (see sect. 4 below) can only be established relative to an assignment of values to the exogenous variables. Since we will be more interested in interventions and counterfactuals than in causal inference, we will simplify things by including the values of the exogenous variables as parts of the model.

[2] The *Markov Condition*, which is formally identical to the Causal Markov Condition, is used in a number of contexts in which a directed graph is not being used represent causal relationships. Thus the modifier 'Causal' is added to the Markov Condition in the context of causal modelling.

[3] In section two above, the concepts 'parent', 'descendant', etc. were defined in terms of structural equations, and represented graphically. In the probabilistic context where we have no equations, the concepts are defined directly in terms of the graph in the obvious way.

[4] If the variables are continuous, the quantification will be over measurable sets of values.

[5] This assumes that all assignments $U_i = u_i$ have positive probability. In the case of continuous variables, the definition of independence is a little bit more complicated.

[6] Halpern and Pearl allow for certain Boolean combinations of assignments of values to variables to count as causes and effects, but we will restrict attention to the simplest case where cause and effect involve only a single variable.

[7] This example is based on one in Hiddleston (2005).

[1] Advocates of this usage of the term 'law' include Cartwright (1999) and Mitchell (1997).

[2] The notion involved here is essentially the same as Woodward's notion of a direct cause (2003), in which direct causes are those causes that can bring about effects while holding all other factors fixed. Directness is relativized to a particular decomposition of a system into parts.

[3] This view is not really at odds with Woodward's, because the path diagrams Woodward uses to represent causal structure are naturally interpreted as diagrams of mechanisms.

[1] In a closely related context (see sect. 2.1 below) David Lewis (1986*a*: 40–1) puts the point like this: 'Careful readers have thought they could make sense of stories of time travel ... ; hard-headed psychical researchers have believed in precognition; speculative physicists have given serious consideration to tachyons, advanced potentials, and cosmological models with closed timelike curves. ... It will not do to declare [these phenomena] impossible *a priori*.'

[2] As Michael Dummett (1954: 28) puts it: 'Why should we lay down temporal precedence as a defining property of a cause? If we can observe that an event of a certain kind is a sufficient condition of an earlier event of some other kind, it does not seem to matter much whether we choose to call the later event the "cause" of the earlier or not: the question rather is why we should not use this observed regularity as we use those that operate from earlier to later; why, when we do not know whether or not the earlier event has occurred, we should not bring about the later event in order to ensure that the earlier had occurred.'

[3] Russell's main claim was that the evolution of physical systems can be fully described without employing causal vocabulary, but he also calls attention to the time-symmetry of the determination relations one finds in physics.

[4] See Chapter 13 on anti-reductionism for more, and Tooley (1987; 1990) as an example of such a view.

[5] We emphasize that our objection here is not that causation is not directly experienced or that it is not analysable in terms of experiences. Rather, it is that hyperrealism entails that causal facts are underdetermined by *all* available non-causal evidence.

[6] An early proponent of this view was F. Ramsey ([1929] 1931: 146), who says that 'from the situation when we are deliberating seems to ... arise the general difference of cause and effect'.

[7] He later revised the analysis (Lewis 2000), but our remarks here are independent of these details.

[8] In what follows we gloss over a number of details that are irrelevant to our discussion. See Ch. 8, Counterfactual Theories, for more.

[9] See n. 1.

[10] Lewis did not give a precise definition of a miracle, but Frisch (2005*b*: 170–3) persuasively argues, partly by appealing to quotes from Lewis and partly by responding on behalf of Lewis to an objection made by Arntzenius (1990), that Lewis requires a spatio-temporal locality condition on miracle size. We take this for granted in what follows.

[11] Albert argues that PH has the status of an additional physical *law*—a view supported by Callender (2004), for example.

[12] A similar conclusion follows from Frisch's (2005*b*: ch. 8) demonstration that there is no fundamental asymmetry of overdetermination in classical electromagnetism.

[13] We would become fatalists about the *distant* future, presumably, believing that nothing we could do would prevent the final state of which God had informed us. But this would have little or no immediate relevance, except under very unrealistic suppositions about what we would otherwise take ourselves to be able to achieve.

[14] It specifies our own future boundary!

[15] We suppose that we cannot tell which way Death is heading before we make our choice. (We know that he takes mortal form, but not *which* mortal form, so it wouldn't help to have

CCTV footage of all the travellers currently moving in either direction.) We thus avoid the so-called bilking argument, exploiting the loophole identified by Dummett (1964).

[16] This kind of zigzag is discussed by Kutach (2002).

[17] Cf. Horgan (1981). Horgan notes an apparently ineliminable circularity in the two-boxers' attempt to justify two-boxing.

[18] The qualification is needed because a one-boxer may want to argue that ordinary causal intuitions are misleading in Newcomb cases: perhaps the decision to one-box should be seen as *retrocausing* the presence of $1,000,000.

[19] In other words, we don't yet believe either disjunct, but would infer each from the falsity of the other. We have put this in terms of disjunctions rather than material conditionals simply to lessen the need to emphasize that the connectives in question *are* material.

[20] The claim that hypothetical reasoning is more fundamental than counterfactual reasoning —that 'counterfactuals are the price we pay for hypotheticals', as Alison Gopnik (pers. comm.) puts it—is certainly not new. But the present argument suggests a new defence of this view, based on the argument that *only* an approach that grounds counterfactuals on hypotheticals can account for the asymmetry of deliberation.

[21] Note that this case provides an exception to TADD: the agent is presented as believing (a) that either he'll take both boxes or there is *already* $1,000,000 in the first; and (b) that he really has a choice as to whether to take one box or two. More on this below.

[22] See Eells (1981; 1982), Horgan (1981), Horwich (1987: ch. 11), and Price (1986; 1991), for arguments of this kind.

[23] It is compatible with this account that PH might be needed to explain the existence of such asymmetric deliberators, as to explain the existence of very much else of a time-asymmetric nature in the world we observe. But this does not *reduce* the account here suggested to the AKL account. By analogy, PH might be needed to explain the existence of creatures who distinguish between left and right, or between locals and foreigner; but this provides no reason whatsoever to regard PH as part of the truth conditions of such distinctions.

[24] See Lewis (1980). Lewis himself thus counts as a subjectivist—even if surely at the objectivist end of the subjectivist spectrum!—in virtue of taking the Principal Principle to be an analytic element of any satisfactory theory of chance.

[25] Another advantage of this approach, compared to Lewis's own, is that it solves the problem of *transitions*. In Lewis's version the small miracle on which the time-asymmetry depends needs to be displaced somewhat from the antecedent in question—which implies that there will always be events between the miracle and the antecedent that turn out to depend counterfactually on the antecedent (e.g. if the vase had smashed later, it would have fallen through the air earlier). In the hypothetical case no such problem arises: the transition is simply our deliberation itself, which we always think of as issuing in rather depending on the ensuing actions.

[26] See e.g. Pearl (2000), Spirtes, Glymour, and Scheines (2000), and Woodward (2003).

[27] This is not to deny that an agent already equipped with causal concepts will regard her deliberations as causes of the acts to which they give rise, but only that the deliberations themselves *depend* on possession of causal concepts. The former circularity is not vicious, whereas the latter would prevent the proposed genealogy from leaving the ground.

[28] As writers such as Woodward (2001; 2003) have emphasized.

[29] Referring to Woodward's (2003) theory, Weslake (2006: 139) puts the point like this: '[G]iven that any variety of counterfactual meeting the criteria of an intervention will give us a variety of manipulation, why is it only some subset of these that we are interested in? Why shouldn't we abandon counterfactual for counterfactual[*], especially if counterfactual[*] will enable us to cause[*] past events? The answer ... is that we can't, in fact, bring about counterfactual[*] antecedents (at least in all cases we know of)—but this is in part a fact about the sorts of agents we are.'

[1] Two omissions merit special mention. First, there is an enormous literature in social psychology on causal attributions as one type of social inference (see Uleman, Saribay, and Gonzalez 2008 for a recent review). Unfortunately, space considerations prevent any serious examination of that research, though social inference is briefly discussed in sect. 2. Second, there is a small but growing body of experimental research on people's explicit (but untutored) judgements about the 'meaning' of the word 'cause' (e.g. Goldvarg and Johnson-Laird 2001; Wolff 2007). This research is still very much in its infancy, and there is growing evidence that the word 'cause' is linguistically ambiguous.

[2] Many examples of such sequences can be found at the website of Brian Scholl's research group, http://www.yale.edu/perception/Brian/demos/causality.html>, accessed 12 March 2009.

[3] One might wonder how we could determine such a thing, given that 6-month-olds are pre-verbal, and even pre-mobile. All of the cited experiments are so-called 'looking time studies'. The basic idea underlying this experimental design is that infants look longer at things that interest them, and stop looking at things that bore them. Thus, if infants who have repeatedly seen Q subsequently look longer at A rather than B, then those infants must think that A is more different from Q than B is. If Q, A, and B are appropriately matched on all dimensions but one, then the infants are arguably conceptualizing (or perceiving) Q and B as the same on that last dimension, while A is different. There are obvious concerns about looking-time studies, and they are notoriously difficult to interpret. Nonetheless, this experimental method is the best we currently have for understanding the mental life of infants.

[4] This operation severs the corpus callosum: the (large) neural connection between the two hemispheres of the brain. It is most commonly performed as a 'last-resort' attempt to control seizures. Cognitive processes can sometimes be localized in these patients to one hemisphere or the other by presenting information to only one eye at a time.

[5] Early statements of causal model theory did not draw a clean distinction between similarity and categorization judgements—$P(O|A)$ and $P(A|O)$, respectively—largely because the experimental designs often did not separate the two. More recent statements of causal model theory seem to have converged on the version presented here (Danks 2007a).

[6] Specifically, the rats were shown a series of $AX+$ trials (i.e. cue A with cue X, followed by the outcome), then a series of $XY+$ trials. At this point, the rats have a relatively strong association between Y and the outcome. A subsequent series of $A-$ trials (i.e. just cue A, and no outcome) led to a reduced associative strength for Y. The rats' use of the $A-$ trials retrospectively to increase the associative strength of X (to 'explain' the $AX+$ trials) is not novel. The surprising part is that the rats seemingly propagate that change outwards (in some sense) retrospectively to revise their learning from the $XY+$ trials: increased strength for X means less learned strength for Y.

[1] And as Yablo (1997) points out, even an involuntary smirk makes no sense unless understood as caused by appropriate mental states.

[2] Notice, however, that there is at least a prima facie tension between the idea that actions are caused by mental *states* of the agent and the idea that actions are brought about *by the agent*. For discussion of this issue, see Horgan (2007).

[3] This is because psychophysical property identities would entail strict synchronic psychophysical property coextensions that are (at least) nomically necessary. Davidson's argument that there are no psychophysical laws, if it works at all, also shows that there are no strict psychophysical coextensions that are either nomically necessary or necessary in some modally stronger way. Hence if it works at all then it also shows that there are no psychophysical property-identities.

[4] Davidson's anomalous monism actually introduced the distinction between type and token identity theories in recent philosophy of mind—he brought about widespread recognition of token identity as an alternative sort of identity theory.

[5] Although we avoid wrangling with it here, there is a distinction to be made between two sorts of properties or types: properties of events and properties of agents. Talk of mental events as efficacious qua mental can apply to either or both kinds of mental properties. When one says, for example, that the token mental event *John's desire for a beer* is efficacious *qua* mental, one might have in mind either *qua* the property *wanting a beer* or *qua* the property *being a desire for a beer* or both. It is an agent who instantiates mental properties such as *wanting a beer*, but it is a token event (probably a neurological one) that instantiates mental properties such as *being a desire for a beer*.

[6] McLaughlin (1989) gave this sort of epiphenomenalism the name 'type epiphenomenalism'. Horgan (1989) coined the phrase 'mental quausation' for causal efficacy of the mental *qua* mental.

[7] We think the Subtraction Argument can also be stated more broadly, so that it applies not only to phenomenal consciousness, but to all mental phenomena.

[8] Chalmers considers several responses to this charge of epiphenomenalism, and may even be willing to accept this as a consequence of his view, though reluctantly (1996: 150–61). Also, for a more detailed version of the argument we present here, see Horgan (1987).

[9] Truth of counterfactuals in vacuous cases (that is, cases in which there is no metaphysically possible world satisfying the antecedent) seems to us to be primarily conventional. Note that although Lewis's (1973*b*: 25) analysis of counterfactuals treats all vacuous counterfactuals as true, he also discusses treating them all as false.

[10] A full-blown counterfactual account of causal relevance is discussed by Le Pore and Loewer (1987). Bennett (2003) gives a lengthy analysis of a similar case without assuming a full counterfactual analysis of causation. It should also be noted that MND is problematic in cases of overdetermination and pre-emption, so one must be very cautious in its use.

[11] We follow Bennett (2003) in dividing premiss (2) into two parts.

[12] We have formulated the argument in a manner intended to remain neutral about the metaphysics of events—e.g. about whether an event is a concrete particular, or instead is an ontologically 'mixed' entity consisting of an object instantiating a property at a time.

[13] Note that since there may only rarely be complete sufficiency in an indeterministic

world, one needs to stretch this definition to allow for quantum indeterminacy.

[14] There are a few other closely related putative reasons for holding propositional attitudes to be extrinsically individuated. For example: (1) the content of your belief depends on causal-historical relations that not every internally indiscernible duplicate shares with you; (2) the content of your beliefs often depends on beliefs of experts who are not present in the context of an internally indiscernible duplicate; (3) a mental state may also be extrinsically individuated due to the propositional attitude, rather than the content of the proposition. For example, your knowledge states may depend on the non-existence of defeaters that you have not even dreamed of; e.g. see Yablo (2003: n. 2).

[15] For a representative sample of strongly externalist theories (i.e. theories asserting that virtually all intentional mental content is extrinsically individuated), see Stich and Warfield (1994). For some treatments of mental intentionality that resist strong externalism, see e.g. McGinn (1989; 1991); Siewert (1998); Horgan and Tienson (2002); Loar (2003); and Horgan, Tienson, and Graham (2004).

[16] Is this a form of functionalism? That depends on how one uses the term. The view does claim that mental *concepts* are functionally definable; but it also claims that these concepts and the mentalistic words that express them denote first-order properties rather than functional properties. Traditionally, the term 'functionalism' was reserved for theories claiming that mental properties are identical to functional properties; under that usage, Lewis's position is not a form of functionalism. But lately, views like his are often called 'filler functionalism', whereas the position originally called functionalism is now often called 'role functionalism'.

[17] On Lewis's theory, multiple realization might better be called *multiple reference*. It's not that one mental property is multiply realized by various different physical properties; rather, one mental concept non-rigidly *refers* to different physical properties, relative to different creature-kinds.

[18] See Horgan (2001*b*).

[19] If some philosopher or scientist could provide decisive non-question-begging reasons for the claim that strong multiple realizability is neither a genuine physical possibility nor even a genuine metaphysical possibility, then a credible theory of mind would no longer need to accommodate the epistemic possibilities lately mentioned; indeed, presumably those erstwhile epistemic possibilities would dissolve, in the face of such reasons. (Likewise, *mutatis mutandis*, for ordinary cross-species multiple realizability.) But providing such reasons is a very tall order indeed.

[20] This section is adapted, with some modifications and additions, from Horgan (1998: sect. 3 and n. 8).

[21] Kim (2005) currently accepts a functional reduction of intentional mental properties, at least he accepts this conditionally on the existence of mental causation at all, and does not accept it for phenomenal mental properties.

[22] For an early version of the idea that one can combine Kim's view of events as property-instantiations with the idea that the constitutive property of a mental event is not a mental property but rather a role-filling physical property, see Horgan (1980).

[23] A notable recent exception is Menzies (2003), a paper in which Menzies applies to

philosophy of mind the contextualist approach to causation that he had already developed on independent grounds.

[24] Many other arguments have been given for contextualism about causation. For example, see Hitchcock (1996); Menzies (2007); Schaffer (2005).

[25] See Holland (1986).

[26] Hartry Field, Fall 1997 seminar on causation at New York University.

[27] For further discussion of this argument for the context-sensitivity of causation, see Maslen (2004*a*: 348–9), and Carroll (2003).

[28] See Chisholm (1955); Goodman (1983); Lewis (1973*b*); Stalnaker (1968).

[29] It has sometimes been objected that the context sensitivity of counterfactuals in these cases arises only from the context sensitivity of the antecedent. So it could be said that in the context of the second assertion the conversationally understood antecedent is 'CM's ring is gold *and still it is non-malleable*.' However, even if this were so, 'CM's ring is gold' is not context sensitive when it occurs alone but only when it occurs within a counterfactual. One does then have a choice between saying that it is the *antecedents* of the counterfactuals that are context-sensitive, or in saying that the *counterfactuals* are context sensitive. But this is a choice without a real difference.

[30] However, both Menzies and Schaffer feature Unger's type of context sensitivity together with a broader context sensitivity. See Menzies (2004) and Schaffer (2005).

[31] Here we assume that the following distinctions roughly coincide: the cause/a cause, the decisive cause/lesser causal factors and the cause/ background conditions.

[32] 'Nothing can better show the absence of any scientific ground for the distinction between the cause of a phenomenon and its conditions, than the capricious manner in which we select from among the conditions that which we choose to denominate the cause' (Mill 1872: ch. 5 [§3]). Also see White (1965).

[33] For example, Lewis ([1973*a*] 1986: 162) is 'unconcerned' with such 'invidious discrimination'.

[34] For further development of an account of mental causation employing sets of relevant possible worlds as the contextual parameter see Horgan (1989).

[35] Hitchcock (1996) gives a contrastive probability-raising account. Field (1997) discusses probabilistic and non-probabilistic versions of a contrastive regularity or law-based view. Holland (1986) presents a contrastive counterfactual account. Maslen (2004*a*) and Carroll (2003) support contrastive counterfactual accounts also. Schaffer (2005) presents a general contrastive account.

[36] Note too that on one version of the contrastive brand of contextualism about causal claims, the contextually pertinent contrast-class(es) partially determine the meaning and truth conditions of an ordinary causal claim without being implicitly a part of the *content* of that claim; i.e. the claim is not covertly *about* contrast-parameters, although such parameters do partially fix the claim's meaning and truth conditions.

[37] And the Contrastive approach can be seen as a species of the Relevant Worlds approach: the relevant possible worlds may be just those that differ from the actual world only with respect to a contrast event or set of contrast events.

[38] See Yablo (2004) for a contextualist approach to causation with a parameter of this sort

as well as other parameters. Note that the Relevant Worlds approach and this approach of holding events fixed correspond to two main approaches to counterfactuals: the Stalnaker–Lewis approach and the older 'metalinguistic' or tacit assumption approach originating with Chisholm. It is interesting that in 1973*c*, Lewis argues that the metalinguistic approach of counterfactuals with its notion of cotenability is simply a reformulation of the Stalnaker–Lewis account of counterfactuals.

[39] See Paul (2004) for a view similar to this, in terms of event aspects.

[40] There is less written about these kinds of mechanisms for the case of causation, but see Maslen (2004*b*).

[41] For an extended discussion of the objectivity of causation and a view related to contextualism ('Causal Persepectivalism'), see Price (2007).

[42] See Horgan and Tienson (1990).

[43] Such variation can happen, across a suitable range of possible worlds, because the full supervenience base for the mental property includes more than just the realizing property. In particular, it includes the externalistic relations that figure constitutively in fixing the content of the supervening mental property.

[44] We are working with the assumption that zombies are metaphysically possible here because it follows directly from an assumption of the Subtraction Argument. (That is, it follows from the assumption that the phenomenal is merely nomically supervenient on the physical and not metaphysically supervenient.) It is, however, tendentious whether zombies are even metaphysically possible.

[45] The MND test requires the non-vacuous truth of the counterfactual 'Had the cause lacked property F then the effect would still have occurred,' but this counterfactual is vacuous in the case in which there are no relevant metaphysically possible worlds. Also, note that there is another path that the contextualist may wish to take here. She may contend that there are still metaphysically possible worlds that are relevant—namely, worlds in which both quale $q1$ and its physical basis p 1 are replaced, subsequently making a difference to the behaviour. If such worlds are relevant, then arguably this shows that the phenomenal property does make a difference!

[46] There are exceptions, though. In some contexts, such as a researcher seeking the neurophysiological etiology of pain behaviour, it would be perfectly appropriate to consider worlds that lack phenomenological experience. This sort of case is not the norm, however, and so poses no problem for our point.

[47] The contextualist may go so far as to acknowledge the existence of contexts in which the mental is efficacious *and the physical is not* (though she need not go this far). Call such contextualists 'extreme contextualists'. Notice that extreme contextualists are also rejecting the first premiss of the Causal Exclusion Argument—the Causal Closure of the Physical—in some contexts. (However, they need not reject the *nomological* closure of the physical in these contexts.) See e.g. Menzies (2008).

[48] See Maslen (2004*b* ; 2005).

Parts of this chapter derive from Mele (1992; 1995; 1997*b*; 2003; 2006).

[1] I borrow the term 'causalism' from Wilson (1989). (Wilson is a non-causalist.)

[2] Alternative conceptions of action include an 'internalist' view according to which actions differ experientially from other events in a way that is essentially independent of how, or whether, they are caused (Ginet 1990); a conception of actions as composites of non-actional mental events or states (e.g. intentions) and pertinent non-actional effects (e.g. an arm's rising) (Mill 1961; Searle 1983); and views identifying an action with the causing of a suitable non-actional product by appropriate non-actional mental items (Dretske 1988)—or, instead, by an agent (Bishop 1989; O'Connor 2000).

[3] This view does not identify actions with *non-actional* events caused in the right way. That would be analogous to identifying genuine US dollar coins with pieces of metal that are not genuine US dollar coins and are produced in the right way by the US Treasury Department, and so identifying genuine US dollar coins would be absurd.

[4] A note on my action variable '*A*' is in order. By the end of the 1970s, a lively debate over the question of action-individuation had produced a collection of relatively precise alternatives. Imagine that Don flips the switch, turns on the light, illuminates the room, and—unbeknownst to him—alerts a prowler to his presence (Davidson 1980: 4). How many actions does he perform? Davidson's *coarse-grained* answer is one action 'of which four descriptions have been given' (ibid. 4; also see Anscombe 1963). A *fine-grained* alternative treats *A* and *B* as different actions if, in performing them, the agent exemplifies different act-properties (Goldman 1970). On this view, Don performs at least four actions, since the act-properties at issue are distinct. An agent may exemplify any of these act-properties without exemplifying any of the others. One may even turn on a light in a room without illuminating the room: the light may be painted black. Another alternative, a componential view, represents Don illuminating the room as an action having various components, including (but not limited to) moving his arm, flipping the switch, and the light going on (Ginet 1990; Thalberg 1977; Thomson 1977). Where proponents of the coarse-grained and fine-grained theories find, respectively, a single action under different descriptions and a collection of intimately related actions, advocates of the various componential views locate a larger action having smaller actions among its parts. In this chapter, I proceed in a neutral way regarding the leading contending theories of individuation. Readers may read my action variable '*A*' in accordance their preferred theory of action-individuation. The same goes for the term 'action'.

[5] On this dispute and its tangential connection to causalism, see Mele (2003: ch. 3).

[6] For a third type, see Mele (1992: 207–10).

[7] '*Proximate* cause' may be defined as follows: x is a proximate cause of y if and only if x is a cause of y and there is nothing z such that x is a cause of z and z is a cause of y.

[8] Many compatibilists are causalists about action. For a causalist account of self-control, see Mele (1995).

[9] For discussion of a third alleged problem for causalism—the problem of so-called negative actions (e.g. intentionally not voting in an election)—see Mele (2003: 146–54).

[10] An exception may be made for the time of the big bang (in universes that begin that way) and times very close to that.

[11] So if the occurrence of *x* (at time *t1*) indeterministically causes the occurrence of *y* (at *t2*), then a complete description of the universe at *t1* together with a complete statement of the laws of nature does *not* entail that *y* occurs at *t2*. There was at most a high probability that the occurrence of *x* at *t1* would cause the occurrence of *y* at *t2*.

[12] For relatively recent luck-based worries about libertarian accounts of free action, see Almeida and Bernstein (2003); Haji (1999); Mele (1995: 195–204; 2006: ch. 3); and Strawson (1994).

[13] Proximal decisions and intentions also include decisions and intentions to continue doing something that one is doing and decisions and intentions to start *A*-ing (e.g. start running a mile) straightaway.

[14] Diana toys with the thought that the agent may be blamed for the decision if past free decisions of his had the result, by way of their effect on his character, that there was a significant chance that he would decide contrary to his best judgement. But she quickly realizes that the same worry arises about past free decisions the agent made.

[15] In Beebee and Mele (2002), it is argued that if the standard assumption of necessitarianism about laws of nature in the contemporary free will literature is set aside and replaced by a Humean view of laws, there is no such ensuring even in deterministic universes. However, it also is argued there that Humean compatibilists face a problem about luck much like the one typical libertarians face.

[16] Also see Kane (2002). Readers who baulk at the thought that an agent may *try to choose to A* (Kane 1999: 231, 233–4) may prefer to think in terms of an agent trying to bring it about that he chooses to *A*.

[17] Kane writes: 'The core meaning of "He got lucky", which *is* implied by indeterminism, I suggest, is that "He succeeded *despite the probability or chance of failure"*; and this core meaning does not imply lack of responsibility, if he succeeds' (1999: 233).

[18] As I understand O'Connor, to exercise the *F*-power is freely to exercise the *D*-power. I do not see how exercising the *D*-power in choosing can itself be sufficient for choosing freely, in a sense of 'freely' closely associated with moral responsibility. If there is such a thing as the power of an agent directly to causally determine his choices, insane agents can have that power, and they can exercise it in making unfree, insane choices for which they are not morally responsible.

[19] If the primitive feature Pereboom describes here can be a feature of agents, it can be a feature of insane agents, and they can exercise the power he characterizes in making unfree, insane choices for which they are not morally responsible. So I do not see how all 'activations of this causal power' can be free choices, in a sense of 'free' closely associated with moral responsibility. See the preceding note for a similar point.

[20] I have used the problem of present luck as a device to organize my discussion of the place of causation in event-causal and agent-causal libertarian views. I believe it is a serious problem for libertarians, but I do not believe it is insurmountable. In fact, although I am not a libertarian myself, I develop a libertarian response to the problem in Mele (2006).

[1] For an example of the last type, see Feldman (1997). Sosa (1993) argues against the causal view and embraces a view of the second type.

[2] For a defence of the opposite view, see Davidson ([1971] 1980). Davidson's view is that, if the killing is done by shooting, then the act of killing is identical to the act of shooting. Now, the shooting clearly causes the victim's death; hence the killing causes the victim's death.

[3] Someone like David Lewis would not want to say this, but because he thinks that the child being an orphan is not an event and thus it cannot enter in causal relations (see Lewis 1986).

[4] I am not objecting to the use of the word 'cause' in legal contexts. But, if the word is used in this way in the law, I would say it picks out a different concept from that used in philosophical contexts.

[5] In fact, I believe that it is in general true that, if c causes e, then the absence of c wouldn't have caused e. See Sartorio (2005).

[6] Someone might object that, in Trolley too, the one's death is *actually* required for the five to survive (in the sense that, as things stood, the one had to die for the five to survive). But this objection fails. In Trolley, the one's death isn't a cause of the survival of the five. The proposal is not that anything that is actually required is a means, but that anything that is actually a *cause* is a means.

[7] In the literature on causation, this type of case is usually called a 'switch'. For discussion of switches, see e.g. Rowe (1989), Yablo (2002), and Sartorio (2005).

[1] In citing these authors, I am not suggesting that they would count themselves as empiricists.

[2] Or, if not, then there exists a chain of events that can be temporally and spatially interpolated between the cause event and the effect event each pair of which satisfies these conditions.

[3] Assuming that one has an acceptable non-causal account of temporal priority.

[4] I am not claiming that these reductions are correct, rather that they are examples of the kinds of reductions that scientists have attempted.

[5] There are more and less restrictive versions of physicalism. Some allow only the entities of completed fundamental physics to exist. Others allow entities referred to by well confirmed theories in any of the natural sciences. It is unimportant for present purposes how strict is the version of physicalism advocated.

[6] 'Realization' has traditionally carried strong connotations of moving from a less concrete to a more concrete (or 'real') carrier of a property. Thus, to realize profits from shareholdings is to cash in the shares, i.e. to realize the property in a more tangible form. The 'real' connotations of 'realization' are no doubt a reflection of this.

[1] Notice that I have not included any structure that fixes how the slices at different times relate to each other (sometimes called a 'rigging' for the spacetime); without this, technically speaking there is no spacetime. Rather than try to depict either full Newtonian spacetime's rigging, or the Neo-Newtonian rigging, I have omitted that element of structure to focus on what matters here: the instantaneous character of the mutual gravitational attraction forces.

[2] The *locus classicus* of philosophical discussions of determinism in Newtonian physics (and subsequent theories) is Earman (1986), recently supplemented and updated by (2007). For a non-technical explanation of determinism and its difficulties in various physical theories, see Hoefer (2003).

[3] For illuminating discussion of both the historical and mathematical aspects of this problem, see Norton (1999) and Malament (1995) and references therein.

[4] Causal anomalies like Laraudogoitia's supertask are not affected by the move to Minkowski spacetime.

[5] See Price and Weslake (Ch. 20 above), and references therein, for further discussion of backwards causation and time travel paradoxes. Maudlin (1994) and Healey (Ch. 33 above), discuss the EPR correlations, which have been found to hold between spacelike separated measurement events.

[6] To readers of this volume, apprised of the great difficulties presented by the notion of causation, the most surprising thing may be the very idea of using causation—which has seemed in need of analysis itself, to most philosophers—as the basis for a reductive analyisis of the relatively clear temporal notions of earlier/later, or temporal betweenness. The project has its roots in Einstein's 1905 relativity paper, which argued that an operational meaning must be provided for the notion of simultaneity.

[7] What Einstein (1905) developed was not only SR, but Maxwell–Lorentz electrodynamics *in* the new spacetime structure of SR. That electrodynamics *is* a physical theory, but one in which causal anomalies or novelties are few and minor, so we give it no space here.

[8] I include here Einstein's Equivalence Principle, the General Principle of Relativity, the Principle of General Co-variance, and Mach's Principle.

[9] See Price and Weslake (Ch. 20 above) for further discussion of causal loops.

[10] Einstein's early (1915–18) hope was that the distinction between inertial and non-inertial motion could be grounded fully in relations to distant masses, with g^{ab} determined by those distant masses and hence playing only the role of a mathematical intermediary: the real causality would be matter–matter. This is the idea known as Mach's Principle. Later Einstein largely abandoned Mach's Principle and was happy to impute direct causal interaction between g^{ab} and matter, as in the quote from MTW. For the full story of Einstein's struggles with Mach's Principle, see Hoefer (1994).

[11] Lewis tried (1979) to find physical grounds for maintaining that most backtracking counterfactuals will be false, but it is now widely accepted that his arguments fail. If we want time-asymmetric causation in GR, we will have to put it in by hand.

[12] The time-symmetry and constraint-based problems are not unique to GR, and in particular both have perfectly clear analogues in Maxwell electrodynamics. But this is just more grist for the Russellian mill.

[1] There is however a lively literature in the criminal law field concerning the relation of mental state, conduct, and consequence (see e.g. Brudner 1998). Many modern criminal theorists separate the involvement question from the question concerning how far down the stream of consequences criminal culpability should extend. Contrast Tadros (2005: 156), who conflates conduct, involvement, and attribution of consequences: 'Causal enquiry ... is sensitive both to moral factors and to states of mind of the defendant.'

[2] *Home Office* v. *Dorset Yacht Co*. [1970] Appeal Cases 1004, at pp. 1027–8 (emphasis added). Though this case is mentioned in the 2nd edn. of Hart and Honoré (1985: 198), this significant dictum is not.

Table of Contents

PART I. THE HISTORY OF CAUSATION   35

PART II. STANDARD APPROACHES TO CAUSATION   128

PART III. ALTERNATIVE APPROACHES TO CAUSATION   241

PART IV. THE METAPHYSICS OF CAUSATION   305

PART V. THE EPISTEMOLOGY OF CAUSATION   394

PART VI. CAUSATION IN PHILOSOPHICAL THEORIES   459

PART VII. CAUSATION IN OTHER DISCIPLINES   565