OXFORD

Understanding Counterfactuals, Understanding Causation

Issues in Philosophy and Psychology

EDITED BY Christoph Hoerl,
Teresa McCormack, & Sarah R. Beck

Understanding Counterfactuals, Understanding Causation

Consciousness and Self-Consciousness

This series arises from the activities of the Consciousness and Self-Consciousness Research Centre at Warwick University. The present volume is one of three interdisciplinary volumes growing out of an AHRC Project hosted by the Centre between 2004 and 2008, on 'Causal Understanding: Empirical and Theoretical Foundations for a New Approach'. Much of the thought behind the project on causal understanding was informed by the idea that making progress on questions regarding consciousness and self-consciousness requires, amongst other things, getting clearer about the extent to which causal understanding is implicated in our awareness of and interaction with our environment, as well as in our grasp of our own place in it. This in turn, however, also requires getting clearer about what exactly causal understanding consists in, and what its distinctive features are. The three volumes explore these issues by integrating philosophical work with experimental and theoretical work in developmental psychology, cognitive psychology, and neuropsychology.

Also published in the series:

Time and Memory, edited by Christoph Hoerl and Teresa McCormack
Agency and Self-Awareness, edited by Johannes Roessler and Naomi Eilan
Joint Attention: Communication and Other Minds, edited by Naomi Eilan, Christoph Hoerl, Teresa McCormack, and Johannes Roessler
Perception, Causation, and Objectivity, edited by Johannes Roessler, Hemdat Lerman, and Naomi Eilan
Tool Use and Causal Cognition, edited by Teresa McCormack, Christoph Hoerl, and Stephen Butterfill

Understanding Counterfactuals, Understanding Causation

Issues in Philosophy and Psychology

EDITED BY
Christoph Hoerl,
Teresa McCormack,
and Sarah R. Beck

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS
Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2011

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by SPI Publisher Services, Pondicherry, India
Printed in Great Britain
on acid-free paper by
MPG Books Group, Bodmin and King's Lynn

ISBN 978–0–19–959069–8 (Hbk.)
978–0–19–969513–3 (Pbk.)

1 3 5 7 9 10 8 6 4 2

Contents

Notes on Contributors

Sarah R. Beck studied Psychology and Philosophy at the University of Oxford and gained her PhD in developmental psychology from the University of Birmingham, where she is currently a Senior Lecturer. Her research is concerned with children's and adults' thinking about time and knowledge, with particular interest in the cognitive processes involved in the development of counterfactual thinking.

Patrick Burns is a Postdoctoral Research Fellow at the University of Birmingham working in the field of cognitive and emotional development. His PhD was from Queen's University Belfast. His research interests include children's causal and counterfactual reasoning, the development of counterfactual emotions such as regret, and the development of executive function.

Ruth M.J. Byrne is the Professor of Cognitive Science at Trinity College Dublin, in the School of Psychology and the Institute of Neuroscience. Her publications include *The Rational Imagination: How People Create Alternatives to Reality* (MIT Press, 2005), and *Deduction*, co-authored with Phil Johnson-Laird (Erlbaum, 1991). She is the former Vice-Provost of Trinity College Dublin.

Dorothy Edgington is a Senior Research Fellow at Birkbeck College, University of London, and a Fellow of the British Academy. She was previously Waynflete Professor of Metaphysical Philosophy at the University of Oxford. She is best known for her work on conditionals, including a long survey article, 'On Conditionals', in *Mind* (1995).

Aidan Feeney completed his PhD at the Centre for Thinking and Language in the Department of Psychology at Plymouth University. He taught at Durham University for ten years before returning to Ireland in 2009 where he is currently Senior Lecturer in the School Of Psychology at Queen's University Belfast. Feeney has published over forty papers and chapters on the psychology of thinking. His interests include pragmatics and reasoning, counterfactual thinking, and inductive reasoning in children and adults. He has co-edited (with Evan Heit) *Inductive Reasoning* (Cambridge University Press, 2007) and is currently on the editorial board of the *European Journal of Cognitive Psychology*.

Caren Frosch completed a PhD at Trinity College Dublin that examined cognitive processes in causal and counterfactual thinking. She has since extended her research by examining children's understanding of causality, and how people's everyday conception of causality might affect counterfactual thinking and decisions in the legal domain. She is now Research Fellow in an ESRC-funded project on children's causal learning at Queen's University Belfast.

Simon J. Handley completed his PhD at Cardiff University before taking up a lectureship post in the School of Psychology at the University of Plymouth in 1993. He has remained at Plymouth ever since, where he now holds a chair in Experimental Psychology. Handley has published more than sixty articles on the psychology of reasoning and judgement. His main research interests include conditional reasoning, dual process accounts of thinking, and the development of thinking amongst typical and atypical populations. He is on the editorial board of *Thinking and Reasoning* and *Memory and Cognition*.

CHRISTOPHER HITCHCOCK is Professor of Philosophy at the California Institute of Technology. He has published articles in most of the leading journals of philosophy and in numerous edited collections. He has also published papers in the fields of computer science, law, linguistics, and psychology.

CHRISTOPH HOERL is Associate Professor in Philosophy at the University of Warwick. Between 2004 and 2008, he was co-director (with Teresa McCormack and Johannes Roessler) of the interdisciplinary AHRC Research Project 'Causal Understanding: Empirical and Theoretical Foundations for a New Approach'. With Teresa McCormack and Stephen Butterfill, he is co-editor of *Tool Use and Causal Cognition* (Oxford University Press, forthcoming).

TERESA MCCORMACK is Professor of Developmental Psychology at the School of Psychology, Queen's University Belfast. She was co-director of the AHRC-funded project on Causal Understanding based at the University of Warwick. Her research primarily addresses issues concerning children's temporal and causal cognition. She has published two co-edited interdisciplinary books: *Time and Memory: Perspectives in Philosophy and Psychology* (Oxford University Press, 2001), with Christoph Hoerl, and *Joint Attention and Communication* (Oxford University Press, 2005), with Naomi Eilan, Christoph Hoerl, and Johannes Roessler. A further volume entitled *Tool Use and Causal Cognition*, co-edited with Christoph Hoerl and Stephen Butterfill is forthcoming with Oxford University Press.

DAVID R. MANDEL holds a PhD in Psychology from the University of British Columbia. He is a senior defence scientist at Defence R&D Canada - Toronto, and he is Adjunct Associate Professor of Psychology at the University of Toronto. His research focuses on human thinking, judgement, and decision making. He is the lead editor of *The Psychology of Counterfactual Thinking* (Routledge, 2005) and co-editor of *Neuroscience of Decision Making* (Psychology Press, 2011).

PETER MENZIES is Professor of Philosophy at Macquarie University in Sydney. He is the author of over fifty articles in journals and edited collections, many of which are concerned with the topics of causation and counterfactuals. He is a co-editor (with Helen Beebee and Christopher Hitchcock) of *The Oxford Handbook of Causation* (Oxford University Press, 2009). He is a Fellow of the Australian Academy of Humanities.

JOSEF PERNER received his PhD in Psychology from the University of Toronto. He was Professor in Experimental Psychology at the University of Sussex and is now Professor of Psychology and a member of the Centre for Neurocognitive Research at the University of Salzburg. He is the author of *Understanding the Representational Mind* (MIT Press, 1991) and over 150 articles on cognitive development (theory of mind, executive control, episodic memory, logical reasoning), consciousness (perception versus action), simulation in decision making, and theoretical issues of mental representation and consciousness. He served as President of the European Society for Philosophy and Psychology, is a Fellow of the British Academy, the Center for Advanced Studies in the Behavioral Sciences at Stanford, the Association for Psychological Sciences, and Member of the Academia Europaea.

EVA RAFETSEDER holds a doctoral degree in Psychology from the University of Salzburg and has received advanced training in psychotherapy at the University of Krems. Her research interests focus on cognitive development, especially the development of counterfactual reasoning and regret in children. Currently she is working at the Neuroscience Institute (Department of

Neurology, Paracelsus Medical University, Christian Doppler Clinic) training as a clinical psychologist.

KEVIN J. RIGGS studied philosophy and cognitive science before obtaining a PhD in developmental psychology from the University of Birmingham. His research interests include children's and adults' theory of mind, counterfactual thinking, and inhibitory control. He has lectured at the Universities of Sheffield and Birmingham, and is currently Reader in Psychology at London Metropolitan University.

JOHANNES ROESSLER is Associate Professor in the Department of Philosophy at Warwick University. He has published articles on issues in the philosophy of mind and epistemology, and is co-editor of *Agency and Self-Awareness* (Oxford University Press, 2003) and *Joint Attention: Communication and Other Minds* (Oxford University Press, 2005).

DAVID M. SOBEL received a BA from Swarthmore College in Psychology and Computer Science, and a PhD in Psychology from the University of California at Berkeley. He is currently an Associate Professor in the Department of Cognitive, Linguistic, and Psychological Sciences at Brown University. He studies causal reasoning, pretending, and theory of mind in young children, using behavioral, computational, and neuroscientific methods. His research has been funded by the NSF and NIH. Many of his research ideas come from watching his daughter play.

JAMES WOODWARD is Distinguished Professor in the Department of History and Philosophy of Science at the University of Pittsburgh, and J.O. and Juliette Koepfli Professor of Humanities emeritus, California Institute of Technology. He is the author of *Making Things Happen: A Theory of Causal Explanation* (Oxford University Press, 2003), which won the 2005 Lakatos Prize, and is the incoming President (2011–2012) of the Philosophy of Science Association.

Introduction

Understanding Counterfactuals and Causation

Christoph Hoerl, Teresa McCormack, and Sarah R. Beck

It seems obvious that there is a close connection between our understanding of certain causal claims and our understanding of claims such as the following: 'If a piece of metal had not burst its tyre, Concorde would not have crashed', 'If less violence was shown on television, the amount of violent crime would be lower', or 'If I were to prune this plant in the next few weeks, it would flower next year'. These latter claims exemplify a type often referred to as a *counterfactual conditional*, or *counterfactual*, for short.[1]

To date, the most prominent way in which the idea of a connection between causal and counterfactual claims has figured in philosophy has involved the idea that the meaning of the former can be analysed, at least in part, in terms of the latter. David Lewis (1973a) has put forward what is probably still the best-known example of a theory following this type of approach—i.e. what is often called a *counterfactual theory of causation*. Viewed more generally, counterfactual theories of causation form a category that also encompasses a number of other approaches that have emerged or come to more prominence since, most notably interventionist theories of causation such as the one put forward by James Woodward (2003; see below for discussion).

It is arguable, however, that at least some of the reasons as to why the general idea of connections between causal and counterfactual claims strikes us as plausible have to do with intuitions that are, at least in principle, quite separable from the issues at stake in counterfactual theories of causation. Counterfactual theories of causation (at least as typically conceived) trade on the idea of connections between the two

[1] The three quoted statements are also all conditionals in the subjunctive mood. As we will see below, there is some controversy over how exactly to construe the relationship between the notion of a subjunctive conditional and that of a counterfactual. For instance, on a narrow understanding of the notion of a counterfactual, as advocated by some theorists, the third type of statement we have quoted (a subjunctive conditional about the future) does not display all the features that should be taken to be characteristic of a genuine counterfactual, because it is not clear that its meaning differs from that of the indicative conditional 'If I prune this plant, it will flower'. Other theorists advocate a much broader notion of a counterfactual, on which even statements not in the subjunctive mood, such as 'If I prune this plant, it will flower' can count as counterfactuals.

types of claim on the level of truth conditions. Yet, we arguably also have intuitions about connections between causal and counterfactual claims in quite a different sense—namely empirically informed intuitions about connections between two types of *thinking* we actually engage in: thinking about causal relations and thinking using counterfactuals. For instance, as Woodward points out in his contribution to this volume, it simply seems to be a datum that people find it helpful, in considering complex causal scenarios, to engage in certain sorts of counterfactual thinking. This is not just an interesting fact about our mental lives, but is also reflected in practices that are part of British and American common law, or that inform engineering decisions at NASA. Similarly, it also seems to be a datum that people will spontaneously generate counterfactual thoughts in response to certain kinds of causal outcomes, especially if they were unexpected and distressing. Again, recognition of this fact goes beyond the anecdotal, and informs, for instance, aspects of post-traumatic stress counselling.

The central idea behind the current volume is that the psychological literature on counterfactual thought and its relation to causal thought provides a large, but as yet largely untapped, potential for exploring philosophical questions regarding the nature of causal reasoning in a way that may ultimately also impact on some of the issues at stake in theories of the type exemplified by counterfactual theories of causation. Conversely, philosophical reflection specifically on the nature of causal reasoning and its relation to reasoning with counterfactuals may help shed new light on some of the theoretical issues at stake in psychological studies that aim to probe, e.g. into the development of these reasoning abilities or the psychological capacities that underpin them. Thus, the chapters in this volume take as their starting point the types of intuitions about connections between causal and counterfactual thinking just mentioned, try to sharpen them up and enrich them through empirical means, and offer theoretical accounts as to how these intuitions are best explained.

This introduction cannot address the full range of perspectives from which these issues are explored in the various chapters in this volume. Rather, our aim in what follows is to draw out a small number of key lines of thought or debates that cut across several chapters, and across the divide between philosophy and psychology.

Counterfactual Process Views of Causal Thinking

How might counterfactual thought and causal thought be related? Perhaps the most ambitious general type of line one might take in this area is to try to argue for what Teresa McCormack, Caren Frosch, and Patrick Burns, in their chapter for this volume (this volume, p. 54), call a *counterfactual process view* of causal reasoning. According to such a view, engaging in counterfactual thought is an essential part of the processing involved in making causal judgements, at least in a central range of cases that are critical to a subject's understanding of what it is for one thing to cause another.

One fruitful way of approaching the different contributions to this volume is to think of them as providing materials, conceptual as well as empirical, for challenging

counterfactual process views of causal thinking, or for responding to such challenges. Some apparent challenges to a counterfactual process view of causal thinking emerge as soon as we look at some of the empirical work on causal and counterfactual thought reported in some of the empirical papers in this volume. Here is a small sample.

- When given a vignette detailing a sequence of events with a negative outcome, and then asked to generate suitable counterfactual statements, the counterfactuals that adults generate focus on antecedents that are different from the ones which they would normally judge to be the *causes* of the outcome. Rather, those counterfactuals seem to be focused on antecedents that would have been sufficient to *prevent* the outcome from occurring (Mandel, this volume).
- Given certain temporal cues, children, like adults, reliably interpret a particular physical arrangement as exemplifying a common-cause structure rather than a causal-chain structure, or vice versa. However, when asked counterfactual questions about potential interventions in the system, 5- to 7-year-olds, unlike adults, do not reliably provide answers that are consistent with their choice of causal structure (McCormack, Frosch, and Burns, this volume).
- When adults are asked to rate the probability of a conditional such as 'If car ownership increases, traffic congestion will get worse', which has a natural causal interpretation, there is little evidence that their answers draw on beliefs based on 'undoing' the antecedent. Beliefs based on undoing the antecedent only appear to come into play when adults are asked, e.g. to judge the *causal strength* of the relation between car ownership and traffic congestion (Feeney and Handley, this volume).

What, if any, implications such findings have for the prospects of a counterfactual process view of causal thought depends on a number of questions, such as the following: Is a counterfactual process view of causal thought committed to the idea that people can in fact explicitly articulate the relevant counterfactuals that underlie their causal judgements, or can we make sense of the idea of merely implicit counterfactual reasoning? To what extent is the truth of a counterfactual process view of causal thought compatible with the idea that people's explicit counterfactual judgements diverge, in certain respects, from their causal judgements? To what extent, if any, does a counterfactual process view of causal thought hinge on a notion of counterfactual reasoning according to which such reasoning necessarily involves some form of mental 'undoing'?

The chapters in this volume offer a variety of different views on these questions, some of which we will touch upon below. However, the above list of empirical observations that prompted these questions, as well as the many further findings reported elsewhere in this volume, also invite a more general comment. Anybody who is primarily familiar with the discussion about counterfactual theories of causation in the philosophical literature and then starts to engage with psychological research on

counterfactual reasoning and its connection with causal reasoning is likely to be struck by the sheer diversity of phenomena that are being studied as part of the latter. In particular, it seems clear from this diversity that there might be a real danger of setting things up in the wrong way from the start if we ask what *the* relationship is between causal and counterfactual understanding, as one might be tempted to if influenced by the discussion about counterfactual theories of causation in philosophy. Rather, there might be a multitude of ways in which different kinds or aspects of counterfactual understanding may interact with aspects of our understanding of causal relationships. We will discuss one important way in which this general consideration might be thought to be relevant to some of the chapters in this volume at the end of this introduction.

Philosophical Challenges

Perhaps the simplest version of a counterfactual process view of causal thought that one might think of would be a straightforward psychological counterpart of something like Lewis' (1973a) version of a counterfactual theory of causation. As already mentioned, counterfactual theories of causation are typically intended to give the truth conditions of causal judgements. The underlying motivation here is the thought that we can capture what it means to say that A causes B by stating that, for the judgement 'A causes B' to be true, a certain kind of counterfactual relationship has to obtain between A and B. It is important to note that, even if this thought is along the right lines, counterfactual theories of causation need not be seen to be descriptive of the psychological processes that people go through when making causal judgements. Yet, it is also easy to see how one might try and make a connection between these two issues: If the truth conditions of causal statements are to be given, at least in part, in terms of counterfactuals, it seems plausible to assume that people's reasoning about causal relationships should be sensitive to the obtaining of the relevant counterfactuals. And one very straightforward way in which one might then account for this sensitivity is by assuming that people actually engage in reasoning with counterfactuals when making causal judgements, i.e. by adopting a counterfactual process view of causal thought.

At its most basic, the kind of counterfactual process view we are envisaging here would have it that we arrive at causal judgements of the kind 'A causes B' by evaluating a counterfactual such as 'If A did not occur, B would not occur'. This kind of view is not actually advocated in any of the chapters in this volume, at least at this level of generality and without further qualification. Nevertheless, it serves as a useful model for considering some of the general types of challenges that counterfactual process views of causal thought face.

One class of challenges one might think of here is discussed in detail in Dorothy Edgington's chapter. Her strategy is to look at some philosophical problems that Lewis' counterfactual theory of causation faces, which also threaten to infect a psychological counterpart of it (of the kind sketched above). One of the key issues she raises is that the

way in which a counterfactual is to be interpreted can often itself depend on what we take the causal facts to be. An example she uses is that of a person tossing a coin and another person saying, 'If I had bet on heads, I would have won'. If the coin has in fact landed heads, the counterfactual is unproblematically true on the assumption that the second person's betting or not betting on heads had no causal impact on the outcome. However, it is not obviously true if her bet might have caused the outcome to be different—imagine that the person who tossed the coin is a swindler who can somehow influence which way it lands.

This poses a threat for a theory like Lewis', which tries to provide the truth conditions of causal claims in terms of counterfactuals. The threat is that the theory will be viciously circular, if the truth of the relevant counterfactuals, in turn, depends on that of certain causal claims. Arguably, this threat of circularity does not just affect Lewis' version of a counterfactual theory of causation, but also has an impact on the prospects of a psychological counterpart to Lewis' theory of the type we have been envisaging. More specifically, if Edgington is right, the problem she identifies with Lewis' account undermines the idea that we can give a *reductive* account of the meaning of causal claims in terms of counterfactual ones. And, as such, it also provides an argument against any counterfactual process account of causal thought with similarly reductive ambitions, i.e. any account that tries to base our understanding of causal claims on prior and independent counterfactual reasoning abilities.

Note, however, that there are versions of counterfactual approaches to causation that are not obviously affected by Edgington's argument. These are approaches that admit that there may be no possibility of giving a reductive account of causation in counterfactual terms, but which nevertheless maintain that an illuminating account of the meaning of causal statements can be given that makes essential reference to the holding of certain counterfactuals. Woodward's (2003) variant of an interventionist approach to causation, for instance, tries to account for the meaning of a statement of the type 'A causes B' in terms of the idea of an invariant relationship between A and B that holds under a range of interventions. This is a counterfactual account, in so far as it interprets 'A causes B' in terms of certain counterfactuals about the consequences of A being intervened on. However, the notion of an intervention is itself a causal notion (cf. Woodward, 2003: ch. 3). First of all, to say that A is being intervened on simply is to say that A is being caused to be a certain way (or caused to occur or not to occur). Moreover, for something to count as an intervention in a given causal system, in the sense relevant to interventionism, it must also meet a set of criteria regarding its own causal *in*dependence from other elements of the system at issue. For instance, we might observe an invariant relationship between A and B, even in the absence of A causing B, if A has a cause that also itself causes B, independently of causing A. In that case, bringing A about by means of this cause doesn't qualify as an instance of intervening on A in the sense at stake in interventionist approaches to causation. (The issue here is basically the one that is also behind the problem of confounding variables in empirical experiments.)

Even though the notion of an intervention is thus itself a causal notion, ruling out a reductive account of the meaning of causal claims in terms of interventionist counterfactuals, interventionists such as Woodward argue that this does not make interventionism viciously circular (see also Woodward, this volume, p. 34). Note, in particular, that we can specify the causal criteria an event must meet in order to count as an intervention that may settle whether 'A causes B' is true without touching on the particular causal relation, if any, that obtains between A and B itself. Thus, the most obviously damaging kind of circularity is avoided.

Suppose, then, that interventionist versions of a counterfactual theory of causation, and, by extension, their psychological counterparts, can avoid the kind of threat of vicious circularity that Edgington identifies in Lewis' theory. The points Edgington makes may still bear on the idea of a counterfactual process account of causal thought in a more subtle way. For they might be seen to put some pressure on the defender of such a theory to make more precise exactly how we should think of the role that counterfactual reasoning has in causal thought. Johannes Roessler, for instance, in his contribution to this volume, contrasts two quite different ways in which one might link up causal reasoning abilities with counterfactual reasoning abilities. According to one version, some counterfactual reasoning ability is required for causal thought, because it is required to grasp some of the essential *commitments* of causal claims. According to another version, causal reasoning is also required to marshal the canonical *evidence* for such claims.

If a reductive counterfactual theory of causation such as Lewis' was correct, this might perhaps also help make plausible the latter version of a counterfactual process account of causal thought. That is to say, if the meaning of causal statements could be reductively analysed in terms of counterfactuals, establishing whether the relevant counterfactuals obtain would arguably constitute the canonical way of finding out about the truth of causal claims.[2] Once we give up the idea of a reductive relationship between causality and counterfactuals, by contrast, this version of a counterfactual process account of causal thought also becomes harder to sustain. The prospects of the alternative version, according to which causal thought involves an ability for counterfactual reasoning because counterfactual reasoning is required to grasp some of the commitments of causal claims, are discussed in detail in Roessler's chapter. In particular, he discusses the extent to which it might be compatible with what he calls 'naïve realism concerning mechanical transactions', which involves the idea that perception can provide us with non-inferential knowledge of mechanical transactions (as opposed, e.g. to mere patterns of movement).

[2] Admittedly, there is scope for further debate on this matter. See, e.g. Woodward, this volume, p. 36, on related matters.

The Developmental Challenge

We have looked at a challenge to counterfactual process views of causal thought that is informed by philosophical considerations, in particular about the meaning of causal and counterfactual claims. But there are also a number of empirical challenges that counterfactual process views of causal thought face. Some of the contributions to this volume by developmental psychologists set out a very basic such challenge: In both verbal and non-verbal tasks, children seem to demonstrate an understanding of causal relations long before they appear to be fully competent with counterfactual reasoning (at least of certain kinds). Thus, it appears that a counterfactual reasoning ability cannot be an essential ingredient in the ability to make causal judgements, if we think of the latter as what is demonstrated in the relevant verbal and non-verbal tasks at issue.

In a very influential 1996 paper, Harris, German, and Mills claimed to have demonstrated that 'young children, including 3-year-olds, can consider counterfactual scenarios in trying to figure out both what has caused a particular outcome and how it might have been prevented' (Harris et al. 1996: 249). Harris et al. explicitly framed their paper in terms of a defence of what we have called a counterfactual process view of causal thought, and took themselves to have found evidence supporting such a view in the way in which children answered counterfactual questions regarding a number of different causal scenarios presented to them in stories. By contrast, the papers in the current volume by McCormack et al., Beck et al., and Perner and Rafetseder all come to a different conclusion. What emerges from them is a picture of counterfactual thought as a very sophisticated cognitive achievement, some elements of which do not in fact develop fully until the age of 10 or 12 years. There are no claims for similarly late developments in causal understanding in the developmental literature.

How is this discrepancy in views to be explained? Those developmentalists who stress the cognitive complexity of counterfactual thought can acknowledge that the children in Harris et al.'s experiments gave correct answers to questions put to them in the form of a subjunctive conditional. However, they are likely to maintain that the children did so on the basis of resources that fall short of genuine counterfactual reasoning, narrowly understood. Thus, for instance, a suggestion that can be found in the chapters by both Perner and Rafetseder and Beck et al., respectively, is that younger children, when asked a question using the subjunctive conditional form 'What would have happened if x had not happened?', actually merely entertain the indicative conditional 'If x doesn't happen, y happens', and answer on that basis. In many cases, at least in the typically rather simple worlds of developmental experiments, y will also in fact be the right answer to the counterfactual question, so the performance of children in counterfactual tasks may mask the fact that they do not genuinely engage in reasoning with or about counterfactuals.

One key underlying thought here is that counterfactual thought is psychologically demanding in as far as it requires, for instance, holding in mind both what could have happened and what actually happened (an idea also explored, within the context of

adult cognition, in Ruth Byrne's chapter; though see Woodward for a critical perspective). This has to be distinguished, the thought goes, from a more primitive ability to imagine what is in fact a non-actual state of affairs, but in a way that falls short of genuine counterfactual thinking. Thus, when asked the counterfactual 'What would have happened if x had not happened?' children might simply draw on their general knowledge to construct in their imagination a situation in which x does not happen, and then answer accordingly. However, they might not link this to their knowledge of what actually happened, as is required for genuine counterfactual reasoning, at least on the view at issue here.[3]

A theoretical perspective on the development of counterfactual and causal thought that differs somewhat from the line of thought just sketched is provided in David Sobel's chapter for this volume. Recall that the line of thought presented above had it that the lack of an ability to engage in genuine counterfactual reasoning might be masked in situations in which children can draw on general background knowledge in answering a question that is put to them in the form of a counterfactual. The idea here is that of a domain-general ability which young children lack (i.e. the general ability to engage in genuine counterfactual reasoning), but the lack of which can be masked in certain circumstances. Sobel, by contrast, can be seen to be pressing the point that, conversely, an existing general ability to engage in counterfactual reasoning might sometimes be masked by the fact that children do not have sufficient knowledge within a domain that they could bring to bear in evaluating counterfactuals about that domain (see also Woodward, this volume). In one of the studies reported by Sobel, for instance, 3- and 4-year-olds were asked counterfactual questions after listening to two stories that were arguably structurally identical. When told a story in which an event fulfils a character's desire, leaving him happy, the children could reliably judge how the character would feel had the desire been left unfulfilled. Yet, when presented with a story in which a character doesn't know that a certain event will happen, and is surprised when it does, the same children could not reliably judge how the character would feel if he had known about the event. As Sobel argues, the most natural interpretation of this finding is that it is to be explained in terms of differences in children's domain-specific knowledge: by the age of 3 or 4, they have already grasped certain facts about the functional role of desires, but still lack a proper understanding of the functional role of knowledge.

[3] There is an influential idea in the literature on children's developing understanding of the notion of a representation that can be seen to provide a historic model for this position. It is a well-established finding that children can engage in pretend play (e.g. acting as if a banana was a telephone) before they can pass false belief tests (i.e. correctly predict the actions of a person who lacks key pieces of information). Perner (1991) explains this developmental dissociation in terms of the idea that pretence only involves the ability to switch between two representations (representing the banana as a banana, and representing it as a telephone), whereas an understanding of false belief requires modelling the other's belief, but as a belief that actually aims at the world one's own beliefs represent to be different. In other words, false belief understanding does not just require entertaining two representations, but relating them to one another.

The advocate of domain-general changes in children's counterfactual reasoning abilities can, of course, admit that the ability to engage in counterfactual reasoning can be constrained by a lack of background knowledge within a domain in the way envisaged by Sobel. What is ultimately at stake in his or her position is the question as to whether, in addition, we can also make sense of, and give empirical substance to, the idea of changes in children's very understanding of possibility. Perner and Rafetseder pursue this issue, for instance, by trying to find ways of disentangling empirically counterfactual and other types of conditional reasoning, and they do find that younger children struggle with cases in which using the latter will not yield the correct answer.

One interesting possibility, however, which emerges from both the chapter by McCormack et al. and that by Beck et al. is that the strongest empirical evidence relating to developments in children's understanding of possibility might in fact emerge from work in which the children are asked to produce counterfactuals that are *different from* the counterfactuals that philosophers putting forward a counterfactual theory of causation have traditionally focused on. (Compare, for instance, McCormack et al.'s discussion of studies on children's comprehension of counterfactuals in situations featuring cue competition, or Beck et al.'s discussion of studies on what they call 'open counterfactuals' and counterfactual emotions.) Clearly, once we look at counterfactuals that are different from the ones that, according to a counterfactual theory of causation, encapsulate the causal relations obtaining in the relevant situation, the particular methodological worry we described in connection with Sobel's contribution to this volume no longer applies. By the same token, however, it might be argued that any developmental differences in children's understanding of possibility that might be found in such studies are of less obvious relevance to the question as to whether some form of counterfactual processing view of causal thought can be sustained. It is to a version of this issue that we turn next.

Two Notions of 'Counterfactual'

The kind of developmental claim that we considered in the previous section—to the effect that genuine counterfactual thought emerges later in development than causal thought—typically hinges on a specific understanding of what a counterfactual is, which we might call a 'narrow' understanding of the notion of a counterfactual. It is important to note here that the issue as to whether young children can engage in genuine counterfactual reasoning, as e.g. Beck et al. or Perner and Rafetseder see it, isn't one about children's linguistic competence. They can allow that counterfactual reasoning abilities might be manifested in tasks that don't require children to produce or evaluate explicit statements of the form 'If x hadn't happened, y would have happened' or similar, but instead look, e.g. at the development of feelings of regret, or at children's understanding of 'almost happened' statements. Rather, central to the 'narrow' understanding of the notion of a counterfactual those authors invoke is the idea of a sharp distinction between counterfactual and indicative conditionals.

A similar narrow understanding of the notion of a counterfactual can be found in Lewis' book *Counterfactuals* (Lewis, 1973b). In Lewis, the idea that there is a sharp distinction between counterfactual and indicative conditionals comes out, for instance, when he explains why 'Subjunctive Conditionals' would not have served as an equally good title for his book. Lewis admits that counterfactuals, as he understands them, are typically expressed in the subjunctive mood, but then he goes on to say that 'there are subjunctive conditionals pertaining to the future, like "If our ground troops entered Laos next year, there would be trouble" that appear to have the truth conditions of indicative conditionals, rather than of the counterfactual conditionals I shall be considering' (Lewis, 1973b: 4).[4]

Not all philosophers working on conditionals share Lewis' views of a sharp distinction between counterfactual and indicative conditionals. Edgington, for instance, makes the point in her chapter that it appears that any acceptable indicative conditional can, as she puts it, 'go counterfactual', given the right context. Broadly speaking, the understanding of the notion of a counterfactual she employs here is that of a past-tense conditional in the subjunctive mood. (Note that this, according to the quotation above, should also count as a counterfactual by Lewis' lights.) Yet, if counterfactuals, in this sense, can be generated by transformation from indicative conditionals, it seems implausible that the original indicative conditional and the resulting counterfactual should require two completely different kinds of semantic analysis. This, in turn, has a direct impact on the prospects for a counterfactual theory of causation as there are obvious examples of indicative conditionals that don't 'track causation', as Edgington puts it, such as 'If she's not at home, she's out for a walk'. If these can be transformed into counterfactuals, as Edgington uses the term, the latter will clearly be unsuitable for figuring in an analysis of causation. Her conclusion is that 'counterfactuals are too wide a class to hope to capture causation in terms of them' (this volume, p. 239).

Woodward, too, thinks that the narrow understanding of the notion of a counterfactual that we have been considering does not single out a phenomenon with a 'fundamentally different type of semantics' (this volume, p. 26) from many other conditionals. However, he develops this idea in a somewhat different way from Edgington. Woodward in fact advocates a broad understanding of the notion of a counterfactual, according to which even a conditional such as 'If I drop this pencil, it will fall to the floor' should be counted as a counterfactual. What governs whether a conditional counts as a counterfactual or not, on this understanding, is whether evaluating it requires 'the insertion of a change into conditions as they are in the actual world, the alteration of some additional features, and the retention of others' (ibid.; see Woodward's chapter for further elaboration of this idea). As Woodward argues, this is the case for the conditional just mentioned.

[4] Lewis (1973b: 3) admits, though, that the title 'Counterfactuals', too, may be misleading, as it might be seen to have the implication, which he rejects, that he is dealing with a class of conditionals the antecedent of which must be false.

There is also a further point that Woodward makes, which is particularly pertinent to the project of the present volume. Even if a narrow understanding of the notion of a counterfactual did turn out to be useful for some purposes, he argues, it might not be the most helpful when it comes to examining potential ways in which causal and counterfactual thought might be connected. Rather, it is much more plausible that it is the broad understanding that we should be focusing on for the specific purpose of examining such connections. In particular, Woodward makes the point that a psychological account of the processes involved in causal thought is likely to assign special significance to causal thought in the context of planning and deliberation. In those contexts, however, conditionals such as 'If I drop this pencil, it will fall to the floor' seem just as central (if not more so) as conditionals such as 'If I had dropped this pencil, it would have fallen to the floor'. Thus, it is natural to think that, if a counterfactual process theory of causal thought is on the right track, the relevant notion of a counterfactual will be the broad notion that Woodward has in mind.[5]

We can look to the chapters by Aidan Feeney and Simon Handley, and by Ruth Byrne, for some empirical work that, whilst perhaps speaking against a counterfactual process view of causal thought on a 'narrow' reading of the notion of a counterfactual, seems consistent with Woodward's views. Indeed Feeney and Handley explicitly interpret one of their results as being in line with Woodward's approach. They used a 'probabilistic truth table task' to study adults' comprehension of what they call causal conditionals, i.e. conditionals most naturally construed as expressing a causal relation. In the task, participants were asked to rate the probability that a causal conditional such as 'If car ownership increases, traffic congestion will get worse' was true, and they were then also asked to rate the probability of each of a set of conjunctions: in each conjunction, the antecedent or a negation of the antecedent was combined with the consequent or a negation of the consequent. Thus, for instance, in addition to the conditional just mentioned, participants would also be asked about the probability of each of the following: 'Car ownership will increase; traffic congestion will get worse', 'Car ownership will increase; traffic congestion will not get worse', 'Car ownership will not increase; traffic congestion will get worse', and 'Car ownership will not increase; traffic congestion will not get worse'. Feeney and Handley found that, in such a task, participants' responses to the original conditional were strongly correlated only with their responses to conjunctions featuring the antecedent, but not with responses to conjunctions featuring the negation of the antecedent. This suggests that they make sense of the relationship expressed primarily by simply imagining a situation in which the antecedent is true, rather than also considering an imagined situation in which the antecedent is not true. Moreover, Feeney and Handley also found the same result when the causal conditional was in the subjunctive mood, suggesting that that there is no sharp distinction in the way the two types of conditionals are understood.

[5] By the same token, any difficulties children may have with certain counterfactuals, more narrowly understood, do not have to stand in the way of such a counterfactual process view of causal thought.

Byrne in fact uses the term 'counterfactual' in a way that is closer to what we have called the 'narrow' understanding. In particular, she takes it that understanding a counterfactual conditional, in contrast to understanding an indicative conditional, involves 'thinking of two possibilities'. This understanding of the term 'counterfactual' differs from that advocated by Woodward, who explicitly rejects a similar idea put forward by Perner and Rafetseder (Woodward, this volume, p. 21f.). Yet, Byrne also claims that when people think about what she calls 'strong causes', they in fact only envisage a single possibility. An example would be thinking about the claim 'Heating water to 100 degrees causes it to boil'. Byrne proposes that people understand this claim by thinking about the possibility that water is heated to 100 degrees and boils; they do not think about the alternative possibility, which is also consistent with the claim, that the water is not heated to 100 degrees and does not boil. Despite the terminological disagreement with Woodward over the term 'counterfactual', Byrne's view is thus actually consistent with his idea that a basic form of causal thought may be centred on the idea that causes are *sufficient* for their effects, which requires a grasp of counterfactuals only in Woodward's broad sense (Woodward, this volume, p. 42).

Causal Judgement and Causal Selection

For Byrne, the idea that understanding causal claims does not always require thinking of two possibilities (and thus a grasp of counterfactuals in the narrow sense she adopts) is connected to the idea of a distinction between 'strong causes' and 'enabling causes'. Enabling causes, she claims, require individuals to think about the same two possibilities as counterfactuals (again, in the narrow sense) do. Thus, there is a specific sort of connection, on her account, between counterfactual reasoning and thought about enabling causes.

The distinction between strong and enabling causes, in Byrne's sense, relates to a topic sometimes referred to as 'causal selection'. Confronted with a scenario in which a certain type of event happens, individuals can make judgements not just as to which factors in the scenario are amongst the causes of the event, which are merely correlated with it because they are other effects of a common cause, and which of them are causally unrelated to it. In our causal judgements we also typically single out one or a small group of factors belonging to the first category as *the* cause or causes of the event in question.

Woodward, in his chapter, argues that the question as to what principles govern causal selection is quite separate from the question as to how we distinguish, for instance, between causation and mere correlation. It is specifically the latter question that Woodward's own version of a counterfactual process view of causal thought, involving the broad reading of the notion of a counterfactual, is focused on. In this, Woodward's main theoretical interests (at least in his chapter for this volume) may be seen to mirror in certain respects those of Lewis, who even went as far as denying that the topic of causal selection was of any significant philosophical interest. As Lewis puts the point,

> We sometimes single out one among all the causes of some event and call it 'the' cause, as if there were no others. Or we single out a few as the 'causes', calling the rest mere 'causal factors' or 'causal conditions'. Or we speak of the 'decisive' or 'real' or 'principal' cause. We may select the abnormal or extraordinary causes, or those under human control, or those we deem good or bad, or just those we want to talk about. I have nothing to say about these principles of invidious discrimination (Lewis, 1973a: 556f.).[6]

A rather different attitude towards the issue of causal selection can be found in Peter Menzies' chapter for this volume. In some respects, Menzies' chapter is the one that most closely adheres to the project of providing a counterfactual theory of causation along traditional Lewisian lines. However, he also points out that there is a class of counterexamples to Lewis' original theory, the common theme of which is that the theory over-generates causes. For instance, if the gardener fails to water a plant and it dies, Lewis' theory counts his failure as a cause of the plant's death. Yet it also counts the Queen's failure to come and water the plant instead as a cause in the same way. This counter-intuitive consequence is simply the result of allowing absences to figure as causes at the same time as treating causal selection as not reflecting any differences of genuine philosophical significance.

Menzies traces back the problem of over-generation of causes to a particular feature of Lewis' theory, namely a centring principle that Lewis imposes on his semantics for counterfactuals. As he points out, Lewis' definition of counterfactual dependence, which is to be the basis for the analysis of causation in counterfactual terms, requires the truth of two counterfactuals:

(i) If c were to obtain, e would obtain.
(ii) If c were not to obtain, e would not obtain.

Because of the centring principle Lewis imposes, however, (i) comes out as trivially true if c and e in fact obtain. Menzies, by contrast, argues that (i) should not be regarded as trivially true if c and e obtain—indeed, we can give examples where (i) seems clearly false, even though both c and e obtain (see also Edgington and Woodward's chapters on related points). Thus, once we give up the centring principle, counterfactual dependence becomes a much stronger condition. And, as Menzies goes on to argue, this allows us to avoid the problem of over-generation of causes that besets Lewis' theory.

Setting aside some of the technicalities, Menzies' suggested strengthening of the definition of counterfactual dependence can be seen as trying to capture the intuitive idea that a cause is a disruption to the way things proceed normally. Thus, if c and e in fact both obtain, but circumstances in which c obtains are not normally circumstances in which e also obtains, we do not count c as a cause of e. This is why we count the

[6] Compare also John Stuart Mill: '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, 1846: 198).

gardener's failure to water the plant as a cause of its death, but not the Queen's failure to do so.

It is at this point that Menzies makes a connection between his own proposal and some of the psychological literature dealing with what is commonly referred to as counterfactual availability. Psychologists have discovered a number of factors that determine which particular counterfactuals individuals are most likely to generate in response to a given causal scenario. It has long been assumed in much of this literature that counterfactual availability might hold the key to causal selection, i.e. to the question as to which factors people single out as the cause (or causes) of a given event. As David Mandel discusses in his chapter, however, the most straightforward way in which one may try to make the connection is not supported by empirical research. When people are asked to generate 'but for' counterfactuals about a certain causal scenario (i.e. counterfactuals corresponding to clause (ii) in the above definition of counterfactual dependence, taken in isolation), their answers typically do not correspond to the answers they would give if asked about the causes of the outcome of the scenario. Even if this is true, though, Menzies' chapter suggests an alternative way in which counterfactual availability might still be related to causal selection. Put crudely, the proposal would be that counterfactual availability governs the way in which we decide whether the definition of counterfactual dependence *as a whole* is met in a particular case or not. In elaborating a proposal along these lines in more detail, Menzies introduces the further technical notions of a deviant and a default counterfactual, before showing how the proposal might be brought to bear on the cases that prove problematic for Lewis' theory.

That there is, in fact, a connection between counterfactual availability and causal selection is also supported by a number of empirical studies reported in Christopher Hitchcock's contribution to this volume. In these studies, a range of factors that influence counterfactual availability are also shown to influence causal selection. What Hitchcock's chapter brings out in particular is that causal selection can be influenced not just by empirical norms, but also by social, legal, and even moral norms. This, though, raises an important general question that might be brought out by noting something of a difference in approach between Menzies and Woodward.

Menzies offers a unified account of the truth conditions of causal claims that effectively builds the normative criteria governing causal selection into those truth conditions. That is to say, on Menzies' account, the very meaning of a causal claim can turn on what we take to be the norms in operation in a particular situation. As a result, Menzies' theory has the feature—which he acknowledges to be controversial—of making the meaning of causal claims in some sense subjective.

Woodward, by contrast, explicitly mentions the fact that causal selection may turn out to be an irreducibly subjective matter as a reason for separating out what he sees as two quite different projects. One of them is to account for causal selection, the other to account for the principles according to which we distinguish between, say, causation and mere correlation. Thus, a key question that Woodward can be seen to be driving at

here is this: To what extent is it possible, by separating out different roles that counterfactual reasoning might play in our thinking about causation, to isolate something like an objective core to our thinking about causation? Or does adopting what we have called a counterfactual process account of causal thought ultimately commit one to a form of anti-realism about our ordinary notion of causation?[7] This is just one way in which considerations about potential psychological connections between counterfactual and causal thought can, ultimately, be seen to lead right back to some of the fundamental types of questions philosophers have been asking about causation.

References

Harris, P.L., German, T., & Mills, P. (1996) 'Children's Use of Counterfactual Thinking in Causal Reasoning', *Cognition* 61: 233–59.

Lewis, D. (1973a) 'Causation', *Journal of Philosophy* 70: 556–67.

—— (1973b) *Counterfactuals*. Oxford: Basil Blackwell.

Mill, J.S. (1846) *A System of Logic*. New York: Harper & Brothers.

Perner, J. (1991) *Understanding the Representational Mind*. Cambridge, MA: MIT Press.

Price, H. & Cory, R. (2007) (eds) *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford: Clarendon Press.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

[7] Related issues of realism vs anti-realism about causation are at the forefront of many of the chapters in Price and Cory (2007). Compare also the way in which the question of realism figures in Roessler's contribution to the present volume.

1

Psychological Studies of Causal and Counterfactual Reasoning

James Woodward

1 Introduction

This paper explores some of the connections between studies of causal and counterfactual reasoning conducted by psychologists and treatments of causation and counterfactuals in the philosophical literature. It is organized as follows. Section 2 compares the different ways in which psychologists and philosophers think about counterfactuals. Section 3 then assesses some claims in the psychological literature about counterfactual reasoning. Sections 4–6 review some philosophical ideas about causation and its connection to counterfactuals, including the distinction between causes and conditions and between actual cause and type cause judgments. Section 7 discusses the 'interventionist account' of causation that I favor, and Sections 8 and 9 discuss the question of whether counterfactual reasoning can be merely 'implicit'. Section 10 discusses the role of causal perception in causal judgment.

2 Counterfactuals: Background

Philosophers reading the psychological literature on counterfactual reasoning are likely to be struck by the very different ways in which they and psychologists use the notion of a 'counterfactual' and cognate notions ('counterfactual thinking', etc.). It is common in the psychological literature to treat conditionals that philosophers regard as counterfactuals (even paradigmatic counterfactuals) as non-counterfactual conditionals of one kind or another and to impose restrictions on what counts as a counterfactual that are not imposed by philosophers.[1] For example, a number of psychologists distinguish

[1] In addition to this more restricted use in connection with conditionals, there is another respect in which usage in psychology departs from usage in philosophy: 'counterfactual' in psychology is commonly used to mean *both* a kind of conditional *and* as an adjective meaning 'false', 'non-actual', or 'containing false premises or assumptions', as in 'counterfactual event', 'counterfactual assumption', and even 'counterfactual syllogism' (Rafetseder, Cristi-Vargas, & Perner, 2010; Beck, Carroll, Brunsdon, & Gryg, 2011)—this is apparently a

sharply between what they call 'future hypotheticals' (roughly claims that *q* will occur if some future condition *p* is met) and what they regard as genuine counterfactuals, while it is common for philosophers to regard many future hypotheticals as perfectly good examples of counterfactuals. More generally, a number of psychologists (including several of the contributors to this volume) require that counterfactuals have one or more of the following features: they must be expressed in the subjunctive rather than the indicative mood, they must have false antecedents, they must have antecedents that are known (by the speaker) to be false or not known to be true, they must not be evaluable just on the basis of default assumptions about the actual world, and/or they must be about particular past events (rather than about the present or future or about relationships between types of events). Many philosophers (myself included) would not accept any of these restrictions.

These differences no doubt derive in part from the very different goals and interests that philosophers and psychologists have in constructing theories of counterfactuals. Psychologists, unsurprisingly, are interested in the role that counterfactuals play, as a matter of empirical fact, in processes of thinking and reasoning. By contrast, many of the philosophers who have been most influential in constructing theories of counterfactuals have approached this subject in the spirit of logicians and/or philosophers of language. They have focused on constructing a semantics of counterfactuals, on providing truth conditions for counterfactuals, and on characterizing the valid inference patterns involving counterfactuals.

I am inclined to think, however, that other considerations besides a difference in goals (or mere non-substantive differences in terminology) are at work. One such consideration seems rooted in an empirical puzzle that is the subject of a number of the papers on counterfactual thinking in this volume. Put very abstractly, the puzzle is that there is evidence that young children (e.g. 3-year-olds) do well on certain tasks involving conditionals but not others (although I note for future reference that there is considerable disagreement about the relevant empirical facts). Somewhat older children (4-year-olds, or perhaps even older children) do much better on both sets of tasks. While psychologists who favor restricted conceptions of counterfactuals and counterfactual reasoning advance arguments (based on claims about the task demands associated with various kinds of conditional reasoning) for these restrictions, an outsider like me is struck by the fact that the conditionals labeled 'counterfactuals' by the restrictionists are pretty much those that (it is supposed) this younger group does not succeed with and that the older group does succeed with. In other words, a characterization of 'counterfactual' is adopted that permits one to say that the younger group is

syllogism with one or more false premises. At the risk of sounding even more curmudgeonly than I do elsewhere in this section, I have to say that I find this usage confusing. There are obvious reasons for not using the same word for a kind of conditional, which can be either true or false, and as a synonym for 'false', especially since there are perfectly good words, including 'false' itself, that more accurately convey what is meant in the latter case.

not and the older group is capable of counterfactual reasoning, hence that a crucial developmental transition occurs with respect to the capacity for counterfactual thinking between the ages of 3 and 4. A more expansive notion of counterfactual thinking, of the sort adopted by many philosophers, would not allow for this description, since subjects in the younger group succeed with some conditionals that involve counterfactuals in this broader sense. For example, in the experiments by Harris described below, 3-year-old children are reported as succeeding in reasoning tasks with conditionals that many philosophers would regard as paradigmatic counterfactuals (e.g. if *X* had taken off her muddy boots before walking on the floor, the floor would have been clean), while some psychologists, adopting a more restricted conception of what is required for counterfactual reasoning, deny that these results are evidence for true counterfactual reasoning. Other tasks involving conditionals on which 3-year-olds fail are said, by contrast, to involve true counterfactual reasoning, so that the upshot is that these younger children can be described as not yet capable of counterfactual reasoning.

I do not mean to suggest that the more restricted usage favored by some psychologists is necessarily illegitimate. If there really is a fundamental difference in the reasoning processes which 3-year-olds, in comparison with 4-year-olds, are capable of engaging in, why not (at least for the purposes of developmental psychology) tailor the notion of a 'counterfactual' to reflect this difference? However, there is a potential problem associated with this path, which is that other possible but more mundane explanations for the differential success rates may be overlooked. For example, in comparing the reasoning tasks on which Harris' children succeed with other reasoning tasks (such as those discussed by Beck, Riggs, & Burns, and by Perner and Rafetseder, this volume) on which they fail, it is worth noting that Harris' tasks involve reasoning about a subject matter (muddy boots and dirty floors) with which the children are likely to have prior experience and familiarity, while the other sets of tasks involve less familiar subject matters, and (in some cases) arguably more complicated antecedents. It is natural to wonder, whether, rather than describing the difference in the experimental results as showing that 3-year-olds lack some general capacity for counterfactual reasoning which is then acquired by 4-year-olds, a better description might be to retain a broader understanding of counterfactual reasoning and describe the results as showing that children are better at tasks involving counterfactuals having to do with familiar subjects than at tasks involving counterfactuals with unfamiliar subjects (or better at tasks requiring counterfactual reasoning involving less rather than more working memory or that impose weaker demands on executive functioning). Presumably, if children are better at conditional reasoning tasks involving simple and familiar subject matters as opposed to complex and unfamiliar subject matters this in itself does not show that the two sets of tasks involve 'different kinds of conditional reasoning' (Perner and Rafetseder, this volume, p. 98).

Put slightly differently, in characterizing the difference between the two sets of reasoning tasks in terms of the idea that one involves reasoning with conditionals that are genuine counterfactuals and the other does not, and accompanying this by an

account according to which genuinely counterfactual reasoning must have feature *X* (where *X* = in the subjunctive mood, having a false antecedent, etc.) while other sorts of reasoning with non-counterfactual conditionals does not have *X*, researchers seem to commit themselves to at least two claims. One is that feature *X* really does distinguish the two classes of conditionals. The other (which seems more in the nature of an implicit assumption) is that it is a very general inability to do reasoning with feature *X* that accounts for why one of the groups fails on the reasoning task, rather than say, subject matter specific features of the task (in which case one might expect that children will succeed with some tasks involving *X* and not others). The analysis in terms of a general inability to engage in tasks involving reasoning with certain sorts of conditionals will be unsupported if either feature *X* fails to distinguish the conditionals on which children succeed from those on which they fail or, even if *X* does distinguish the two sets of conditionals, there is no evidence supporting the claim that it is this difference which explains the pattern of reasoning success and failure with the conditionals. Suppose, for example, that 3-year-olds do better at answering (some) verbal questions concerning conditionals about the future (where it is not known whether the antecedents of these conditionals are true or false) than they do at answering questions concerning conditionals about particular events in the past with antecedents that are known to be false. It is a further hypothesis, requiring additional evidence, that the *cause* of this difference in performance is that the two conditionals differ in the way just described and that qualitatively different kinds of reasoning are involved in the evaluation of the two sorts of conditionals. One disadvantage of labeling the first set of conditionals 'future hypotheticals' and the second set 'counterfactuals' is that it may encourage researchers to simply assume that it is this difference between the two sets of conditionals that explains the differential performance and that the differential performance is a general, domain-independent feature of any pair of tasks involving the two sets of conditionals. Of course (to repeat) this claim *may* be true as an empirical hypothesis, but it cannot be established just by noting that children perform differently on some pairs of tasks involving the two sets of conditionals.

There is a second consideration that also bears on the issue of how narrowly or broadly to construe the notion of a 'counterfactual'. In both philosophy and psychology, counterfactuals are of interest in part because of the way in which (it is thought) they connect up with other concepts and phenomena. For example, many researchers have supposed that there are intimate connections between causal claims and counterfactuals and many researchers have thought that the ability to reason with counterfactuals is closely bound up with planning and deliberation (because to do these I need to consider, e.g. counterfactuals having to do with what would happen if I were to do action *A*, action *B*, etc.). However, to the extent that such connections exist, the 'counterfactuals' that are connected to causation, planning, and so on, are almost certainly conditionals that are counterfactuals in the broad sense favored by philosophers (the sense that counts many future hypotheticals as counterfactuals) rather than in the narrower sense favored by psychologists. Thus, if one restricts oneself to the

narrower notion of counterfactual favored by many psychologists, such counterfactuals are likely to have at best a rather limited role (in comparison with other conditionals) in planning, in elucidating causal notions, and so on. Put somewhat differently, if the philosophers who talk about a connection between causation and counterfactuals are using 'counterfactual' in a broad sense and the psychologists who ask whether there is such a connection are assuming a much narrower notion of counterfactual, there will be a great deal of mutual talking past one another.

In order to further explore some of the issues just described, I begin with a brief account of the notion of a counterfactual conditional, as this is understood by philosophers.[2] Although there is some disagreement about which conditionals count as counterfactuals, I think it fair to say there is a general philosophical consensus that what makes a conditional a counterfactual has to do with its semantics: with the meaning or content or truth conditions of the claim in question and the valid reasoning patterns in which it figures and *not* with the grammatical or syntactic form or mood in which it is expressed. Moreover, there is fairly widespread acceptance among philosophers of the idea that the semantics of counterfactuals can be usefully elucidated in terms of possible worlds. 'If p were the case, then q would be the case' means something like: in those possible worlds which are closest or most similar to the actual world in which p is true, q is also true. To illustrate, suppose that the counterfactual of interest (obviously, I'm using 'counterfactual' here in the broad, philosophical sense), uttered in reference to a dry, well-made match, in background conditions which are optimal for ignition (absence of wind, etc.) is:

If I were to strike this match, it would ignite.

Assume that this counterfactual is true. Assume also that I never strike the match, and it never ignites so that the antecedent as well as the consequent of (1) is false. In evaluating (1) within the possible worlds framework, I consider a situation or 'world' in which the antecedent is true (I do strike the match) and imagine this happening in a way that retains or preserves relevant features of the actual world (this is what the notion of a possible world that is 'close' to the actual world is supposed to capture). Thus I retain such features of the actual world as the dryness of the match, the absence

[2] In what follows I focus on the way in which the semantics of counterfactuals figures in philosophical discussion. It is also natural to ask whether (or in what respects) philosophers understand that semantics as having psychological implications. I think that the answer is unclear. Presumably it is possible to regard the semantics as simply a device for representing the truth conditions of counterfactuals and capturing certain inference patterns, with no implication that it has any further psychological reality. On the other hand, a number of philosophers, including both Mackie and less obviously Lewis, often write as though they think that their proposals capture, perhaps at some very abstract level of description, processes that go on or sometimes go on when people evaluate counterfactuals—one imagines the antecedent inserted into a world appropriately similar to the actual world and considers what will or would happen, etc. Presumably doing this does not require that one explicitly and consciously think in terms of possible worlds, a similarity metric over them, and so on—it is enough that this apparatus can be used to describe what subjects do. I will return to this issue of psychological reality in subsequent sections.

of wind, and the laws of chemistry and physics governing match ignition. However, the relevant closest possible world is *not* one in which the match is struck but *every* other feature of the actual world is left unchanged. For one thing, in the actual world the match does not ignite, but for purposes of understanding (1) this should not be a feature of the closest possible world in which the match is struck since in this possible world, (1) is false. Instead, the notion of a closest possible world should be understood in a way that involves a change in some features of the actual world in addition to those required by the truth of the antecedent of the counterfactual, while other such features are left unchanged. (Philosophers and others have various specific proposals about how to determine which features should be retained and which changed in such cases. I will largely put these aside in what follows, but will say that I agree with Dorothy Edgington, this volume, that doing this relies on causal information—roughly what is left unchanged is what is causally independent of the antecedent and what must be changed is what is causally dependent on the antecedent. For this reason alone, attempts to 'reduce' causal claims to claims about counterfactual dependence that do not presuppose causal notions seem hopeless—see Section 7 below).

I have stressed these points because portions of the psychological literature, including some of the papers in this volume, appear to claim that some conditionals such as some future hypotheticals that many philosophers regard as counterfactuals but which psychologists do not regard as counterfactuals should be understood as claims that just have to do with the actual world and that do not involve any consideration of what is 'non-actual', or at least do not involve relating the (non-actual) possibility stipulated in the antecedent of the conditional to the details of the actual world. (It is partly on these grounds that the psychologists claim that the future hypotheticals are not true counterfactuals; moreover the idea that they only involve reasoning about the actual world is supposed to explain why reasoning with them is easier than reasoning with genuine counterfactuals.) I agree that to count as a genuine counterfactual, a conditional should have a content or semantics that conforms to the possible world framework just described, but unlike what appears to be the views of Perner and Rafetseder, and Beck et al. (this volume), I think that so-called future hypotheticals and many of the conditionals involved in what Perner and Rafetseder call 'basic conditional reasoning' do have this possible world semantics, for reasons that I will attempt to explain in more detail below. That is, on my view, many of the conditionals that are labeled future hypotheticals by psychologists are not just about the actual world, but instead connect the possible and the actual just as other counterfactuals do. Hence, whatever the correct explanation for why (to the extent that this is the case) young children do better on tasks involving future hypotheticals than tasks involving, say, counterfactuals concerning past events, the explanation cannot be that the former but not the latter require reference only to the actual world. I also hold, in contrast to Perner and Rafetseder, that there is no useful sense in which, when counterfactuals are understood in terms of the possible world framework just described, they should be understood as requiring two different models, one having to do with what actually

happens and the other with what happens in a merely possible world. Instead we should see the possible world framework as involving a single common model, in which both the possible and the actual can be related. In particular, when we evaluate the truth of a counterfactual with a false antecedent, we (so to speak) insert a condition that is non-actual into a model that in many other respects retains features of the actual world, so that a single model integrating both the actual and the possible is employed. Thus, as noted in connection with (1), supposing that in actuality the match is not stuck, we consider a (single) model or world that, while departing from the actual world in this respect, nevertheless retains such features of the actual world as the dryness of the match.

Note also that there is nothing (or at least nothing obvious) in the possible world understanding of counterfactuals that requires that counterfactuals can be expressed only in the subjunctive mood. Instead, if a conditional expressed in the so-called indicative mood conforms to the semantics just described, then it will count as a counterfactual according to the possible world account. Nor is this an arbitrary and easily dispensable feature of the possible worlds account. One obvious reason for not restricting counterfactuals to conditionals in the subjunctive mood is that it looks as though in many cases conditional claims expressed by subjunctives and claims that are expressed by indicatives differ little if at all in meaning and truth conditions. As an illustration, suppose that I hold a pencil in my hand and in circumstances in which, apart from my hand, there is an unimpeded path from the pencil to the floor. Now consider the following two sentences:

(2) If I were to drop this pencil, it would fall to the floor.

(3) If I drop this pencil, it will fall to the floor.

(2) is expressed in (what is usually taken to be) the subjunctive mood, (3) is an indicative. Nonetheless, (2) and (3) seem very close in meaning and seem to have identical or very nearly identical truth conditions. If what makes a conditional a counterfactual conditional (as opposed to a conditional of some other kind) is its meaning or truth conditions, and if (2) is regarded as a counterfactual, it seems hard to resist the conclusion that (3) should be regarded as a counterfactual claim as well. Since (2) is regarded by many philosophers as a paradigmatic counterfactual, many adopt a similar view about (3). Indeed, as I understand the empirical literature, if young children have mastered the subjunctive constructions at all, they do roughly as well with conditionals like (2) and with conditionals like (3), which again seems to suggest that they may not differ in any fundamental way in terms of the reasoning they involve.

By contrast, as noted above, in the psychological literature (3) is often described as not a true counterfactual but rather as a 'future hypothetical' or as some other form of (non-counterfactual) conditional. Moreover, the very fact that children succeed on tasks involving both conditionals like (2) and conditionals like (3) is taken by some authors (e.g. Perner and Rafetseder, this volume, if I have understood them correctly) to show that even success on tasks involving conditionals like (2) is not good evidence regarding children's ability to reason successfully with counterfactuals, on

the grounds that subjects given tasks involving conditionals like (2) might answer them by exploiting their equivalence to conditionals like (3) which are taken to be non-counterfactuals.

Another important difference between the psychological and philosophical literatures concerns whether there can be counterfactuals with true antecedents. According to the most influential philosophical accounts, such as Lewis', the answer to this question is 'yes'—indeed, according to Lewis, a counterfactual of form 'if *p* were the case, then *q* would be the case' is always, automatically true whenever *p* and *q* are true.[3] This 'always true' feature of Lewis' account is arguably a mistake (and is, moreover, perhaps a dispensable feature of the account), but the more general claim that it is possible for a counterfactual to have a true antecedent (and to be true in at least some circumstances in which it has a true antecedent) is not easily avoidable if anything like the standard semantics for counterfactuals is correct. The reason is this: given a counterfactual of form, if *p* were the case, then *q* would be the case, then, if *p* is true of the actual world @, that world will be among the possible worlds that are closest to the actual world (since @ is identical with the actual world, it can't fail to be among the closest to it) and since by hypothesis *q* is also true in that world, the truth conditions for 'if *p*, then *q*' will be satisfied as long as there are no other worlds equally similar to the actual world in which *p* holds and *q* fails to hold. Presumably this last condition will be met in at least some cases. Moreover, quite apart from issues about the implications of possible world semantics, it seems very natural to regard many counterfactuals with true antecedents as true. As an illustration, consider again the conditional (2), which most philosophers would regard as a counterfactual claim. Suppose that in actual fact I do drop the pencil and it falls to the ground. The antecedent of (2) is thus true and yet this behavior on the part of the pencil seems to be not just compatible with the truth of (2) but particularly good evidence that (2) is true. It seems odd to take the occurrence of such evidence to establish that (2) is not really a counterfactual or not a true counterfactual. The view that a genuine counterfactual cannot have a true antecedent seems to require that, as matters turned out (the pencil being dropped), (2) was not really a counterfactual (it was some other kind of conditional), although (assuming that *some* conditionals of form (2) with false antecedents are genuine counterfactuals), (2) perhaps would have qualified as a counterfactual if the pencil was not ever dropped. Moreover, if, as will commonly be the case, at the time at which (2) is uttered, the speaker or her audience doesn't know whether the pencil will be dropped, then (again assuming that sentences of form (2) with false antecedents do count as counterfactuals) that person

[3] This is a consequence of Lewis' adoption of what he calls the 'strong centering principle', according to which no other world is as similar to the actual world as the actual world is to itself. To see that this principle is probably mistaken when construed as an account of our practices of evaluating counterfactuals consider a scenario in which an indeterministic coin which has probability one half is tossed at time *t* and lands heads. The counterfactual, 'if this coin were tossed at time *t*, it would land heads' has a true antecedent and consequent but is regarded by most people as false. For additional discussion of the status of the strong centering principle, see List and Menzies (2009).

will not yet know whether or not (2) is a genuine counterfactual—he or she will have to wait until it is learned whether the pencil is dropped before knowing he or she is dealing with a counterfactual or some other kind of conditional. From this point of view, making whether or not (2) is a counterfactual turn on whether or not its antecedent is ever realized (or on whether it is known by the speaker that it is realized) seems to fit badly with how we go about reasoning with and testing counterfactual claims.[4]

Consider in this connection Beck et al.'s (this volume, p. 114) claim that 'The difference between thinking about a counterfactual event and a future hypothetical event is that in the former case what is being imagined is known to be false.' Interpreting this as a claim about conditionals, rather than 'events', it might be taken simply as a stipulative definition according to which to count as a counterfactual a conditional must have a false antecedent. I take it, however, that Beck et al. mean to suggest something stronger—roughly something like 'reasoning with conditionals the antecedents of which are known to be false (that is, counterfactuals in Beck's et al.'s sense) is of a qualitatively different kind or requires qualitatively different abilities from reasoning with otherwise similar conditionals the antecedents of which are not known to be false'.[5] (In particular, Beck et al. think that reasoning with counterfactuals in their sense imposes demands on inhibition and working memory that are not imposed by reasoning with other sorts of conditionals.) I don't claim that this contention is necessarily mistaken, but I think that it is in some tension with the observations in the preceding paragraph. Again, suppose you are reasoning with a future hypothetical like (2) above and don't know whether its antecedent is true. Is it really plausible that you need to switch to a different kind of reasoning ('counterfactual reasoning') if you then learn that I will not drop the pencil? Of course it might be true that reasoning with (2) is *harder* in this case, and perhaps also true that it is harder because it imposes greater inhibitory demands (see Section 3), but presumably this is different from claiming that in this case a qualitatively different kind of thinking is required.

Another point that is worth making about counterfactuals is that (at least as philosophers think about them) they can concern both generic or repeatable relationships between types of events and claims about the relationship between particular individual events. Moreover, each of these alternatives—generic or particular—can concern

[4] Of course this difficulty can be avoided by denying that (2) is a genuine counterfactual, whether or not its antecedent is false. This would be a more coherent position than making whether (2) is a counterfactual depend on whether its antecedent is ever realized.

[5] We see here, incidentally, an illustration of difficulties that can result from using 'counterfactual' both as an adjective applying to events meaning 'non-actual' and as a name for a kind of conditional. If 'counterfactual' just means 'non-actual', then it is a trivial tautology that thinking about a 'counterfactual event' involves imagining 'something known to be false'. Apparently, it is easy to slide from this truism to the idea that counterfactual conditionals must refer to non-actual events in their antecedents (i.e. must have false antecedents), and from this to the idea that there is a fundamental difference between the psychological processes associated with reasoning with counterfactual conditionals and with future hypotheticals. Again, there may as an empirical matter be such a difference, but its existence doesn't follow just from the truism about what is involved in thinking about 'counterfactual events'.

either matters in the future, present, or past and, moreover, generic counterfactual relationships can be timeless in the sense of referring to no particular time. Thus we have, in addition to obvious examples of counterfactuals about particular past events, counterfactuals like

(4) If dinosaurs had had no tails, they would have toppled over. (Generic but about the past.)

and

(5) If kangaroos had no tails, they would topple over. (Generic, timeless—the example with which Lewis begins his book, *Counterfactuals*, but perhaps not a counterfactual at all, according to some psychologists.)

Similarly, while counterfactuals about particular events are often about the past, and counterfactuals about the future often relate types of events, this is by no means always the case:

(6) If I were to drop this pencil which I am now holding in my hand at exactly 12: 45 pm PST, it would fall to the ground. (Particular events, future.)

Again, this stands in contrast to the ways in which at least some psychologists distinguish (what they call) 'counterfactuals' from other sorts of conditionals. For example, at the beginning of their paper, Perner and Rafetseder (this volume, p. 91) write that they will

> focus on *counterfactual reasoning* required by subjunctives about the past and *basic conditional reasoning* for (timeless) conditionals typically expressed in the present to capture recurring regularities.

Here the apparent implication is that there is something distinctively 'counterfactual' about the former ('subjunctives about the past') but not the latter conditionals. By contrast, many philosophers regard conditionals like (5) and (6) as genuine counterfactuals as well.

I said above that one reason that philosophers are inclined to regard conditionals like to as genuine counterfactuals is that they seem to conform to the possible world semantics they take to be characteristic of counterfactuals. To spell this out, consider (2). In understanding/evaluating (2), according to the possible world analysis, I should consider a scenario in which the pen is dropped, various other changes which are required by this assumption are also introduced (the falling pen displaces air molecules, etc.), but in which many other features of the actual world are retained (gravity is still operative, if in the actual world there is no barrier between my hand and the floor, no such barrier is introduced, but if there is such a barrier in the actual world, it is retained, and so on). Assuming that at the time at which (2) is uttered, I do not know whether the pencil will be dropped, I do not know whether features like the release of the pen that are assumed in this scenario will characterize the actual world or not, but in understanding/evaluating (2) I don't need to know this—I proceed in the manner

described above regardless of whether the pen in fact is dropped or not. Thus in both cases (whether or not the pen is in fact dropped), the insertion of a change into conditions as they are in the actual world, the alteration of some additional features, and the retention of others is involved. Similarly for a 'timeless' counterfactual like (5), although here it is presumably known to the speaker that the antecedent is false. Again, one proceeds by considering a non-actual possible world in which kangaroos lack tails but are otherwise similar to actual kangaroos, the force of gravity is as it is in the actual world, and so on. So here too, one does not just consider what is actual but employs some reasoning process of relating the possible to the actual. It is basically for this reason that many philosophers have not tended to think that there is a fundamental difference (at the level of semantics and truth conditions) between conditionals like (3) and (6), and subjunctive conditionals involving particular events in the past. Of course, it may be that reasoning with conditionals like (3) and (6), on the one hand, and subjunctive conditionals about the past, on the other, imposes different sorts of cognitive demands (again, see below), but if the possible world representation is on the right track, the explanation for this difference cannot be that the first set of conditionals only concerns what is actual while the second concerns fitting together the possible and the actual. In other words, the different cognitive demands that may be imposed don't arise because we are dealing with two kinds of conditionals with fundamentally different semantics.

3 Psychological Studies of Counterfactual Reasoning

With this as background, I turn to a more detailed look at some of the experiments involving children's reasoning with conditionals that have been conducted by psychologists. Riggs et al. (1998) conducted an experiment in which children's performance (at ages 3–4 years) on two tasks was compared. In both tasks, pieces of paper were sorted into one of two boxes, depending on whether they had drawings on them or were blank. In one task, which Beck et al. (this volume) describe as having to do with 'future hypotheticals' (rather than counterfactuals) children were asked

(7) 'If I draw on this paper, which box will it go into?

In the second task, the children were asked, regarding some piece of paper that had a drawing on it,

(8) If I had not drawn on the piece of paper, which box would it go into?

The authors think of the conditional associated with this question as a 'counterfactual' rather than as a future hypothetical. Riggs et al. report that the children found it much easier to answer the future hypothetical question than the counterfactual question. In another experiment, Robinson and Beck (2000) presented children with scenarios in which a toy car drives along a road with a garage at either end. Here children found it

easier to answer (what Robinson and Beck take to be) the future hypothetical question 'If next time he drives the other way, where will he end up' than the counterfactual question 'If he had driven the other way, where would he be?'.

One obvious difference between the two sets of questions is that the questions Robinson and Beck label 'counterfactual' concern particular events in the past—in these questions, a past event is known to have occurred (a piece of paper has a drawing on it, a car has been driven to a particular location) and the children are asked what would have happened if this event had not occurred. By contrast, the future hypothetical questions concern what would happen if some event in the future were to occur (and where it is unknown whether or not that event will occur) and this doesn't require imagining the non-occurrence of an event known to have actually happened. Beck et al. suggest that this is why children have more difficulty with the former task; they suggest that in former tasks but not the latter subjects must entertain as true things known to be false and (they take this to be distinct—see below) to resist responding with what they know to be true. By contrast in the case of (what the authors call) 'future hypothetical thinking', the subjects do not know the true state of affairs, so no such negation of the actual is required. Beck et al. claim that some time around the age of 4 children acquire this ability, and that this accounts for the substantial improvement of 4-year-olds over 3-year-olds in answering counterfactual questions. They link this, in turn, to improvements in children's executive function and inhibitory control, as well as working memory, from 3 to 4 years; they claim that improvements in these make it easier for children to inhibit their knowledge of what actually happens, which is essential to assessing counterfactuals but not future hypotheticals.[6]

I do not deny the intuitive appeal of the idea that successfully answering a query like (7) imposes greater inhibitory demands than answering a query like (8). However, I think that more evidence and argument are required to establish that this feature by itself accounts for the difference in children's performance on these tasks. For one thing, as Beck et al. recognize, other researchers have obtained results in which 3-year-old children do perform successfully on tasks involving conditionals about the past with false antecedents. For example, Harris et al. (1996) presented 3-year-olds with scenarios in which, e.g. children observed a doll make muddy footprints on a clean floor and were then asked, what if the doll had taken her shoes off? Would the floor still be dirty? 75 per cent of 3-year-olds answered this question correctly. (Similar results are reported by German and Nichols (2003) and by Sobel (this volume). If the mere fact that a

[6] Beck et al. also distinguish between thinking about what one knows to be false and resisting or inhibiting what one knows to be true. They claim that what they call counterfactual syllogism tasks involve the former but not the latter. (An example of such a syllogism: Children are asked to imagine that all sheep are purple and that Jenny is a sheep, and then asked whether Jenny is purple). They claim, by contrast, that genuine counterfactual reasoning involves both abilities. I find this puzzling. It is hard to see how imagining or thinking that sheep are purple while recognizing that this is not true does not involve some sort of suspension or avoidance of the thought that they are white, which is presumably well known to the children. When a child responds to the syllogism task by saying that Jenny the sheep is purple in what sense is he or she not 'resisting' what he or she knows to be true?

conditional involves the 'negation' of an actually occurring past event creates problems with inhibition for 3-year-olds (so that the possibility assumed in the antecedent of the counterfactual conflicts with what is known to be actually true), why don't we see the same effect in Harris' experiments? A similar point might be raised for other forms of reasoning/cognition apparently successfully employed by young children. For example, most researchers seem to agree that children engage in fantasy and pretend play considerably before they are 3 years old: children pretend that, e.g. a block is a telephone, and act in a way that follows from or is coherent with this assumption, by pretending to talk on the phone, hang up, etc. If young children have a general problem with inhibiting what they know to be actual when they entertain non-actual possibilities, why does this not interfere with their ability to engage in pretend play? After all, there is general agreement that the children understand that the block is not really a telephone.[7] (Just to clarify: This argument does not require the assumption that pretend play and the ability to engage in counterfactual reasoning involve exactly the same abilities; the point is that pretend play seems to involve some form of inhibition of the actual, just as counterfactual reasoning does.)

A second issue concerns what follows if the greater inhibitory demand/working memory account of the difference between 3-year-old and 4-year-old performance on queries like (8) is correct. As noted above, Beck et al. (and other authors as well) seem to move from this account to the more general conclusion that there is a qualitative difference of some kind in the reasoning that is employed in tasks associated with conditionals like (7) and (8). It is not obvious that this is warranted. Suppose I can do multiplication problems involving two digit numbers in my head very successfully but have considerable difficulties with problems involving multiplication of four digit numbers. Suppose that this is due to limitations on my working memory (or limitations on my ability to inhibit distracting stimuli) that affect performance on the latter task. Would it be correct to conclude that the two tasks involve qualitatively different kinds of reasoning or that there are two different forms of multiplication related reasoning only one of which I have mastered? Presumably not. It seems more natural to describe the case as one in which a single form of multiplication reasoning is employed in both tasks but one set of tasks is harder or more demanding than the other, with the level of demandingness being a matter of degree, rather than representing a qualitative difference. Indeed, my working memory/inhibition limitations may well not be specific to multiplication at all. Thus even if it is correct that the greater difficulties younger

[7] Beck et al. also find that when children are given the tasks involving counterfactual syllogisms described in footnote 6 there is no relationship between their performance and measurements of inhibitory control, contrary to what one might expect. Beck et al. attempt to explain this result by contending that such syllogisms are presented in a 'fantasy format' and that this reduces demands for inhibitory control—that is, successful performance is possible without such inhibition. A similar claim might, I suppose, be made about why children succeed at pretend play. The obvious problem with this suggestion is that, in the absence of some further explanation of why fantasy and pretend play have this effect of reducing the need for inhibitory control, it seems completely ad hoc: inhibitory control is required to reason with possibilities that are contrary to fact, except when it isn't.

children experience with what Beck et al. call counterfactuals in comparison with future hypotheticals is due to the greater inhibitory/working memory demands of the former, it isn't obvious that this warrants the conclusion that there is a fundamental difference in the reasoning processes associated with the two kinds of conditionals.

Finally, it is worth noting that if the difficulty in inhibiting the actual account of the children's differential abilities is correct, it seems to make a number of additional predictions that are worth investigating. First, it seems to predict (at least ceteris paribus) that 3-year-olds will also have difficulty with reasoning involving timeless conditionals like (5), since these also involve negating or inhibiting facts that are known to be true.[8] For similar reasons, given subjects with inhibitory limitations, they should have greater difficulties with future hypotheticals like (6) under a condition in which they are convinced (or 'know') that the antecedents of these conditionals would not be realized, than under a condition in which they don't know this.

Turning now to Perner and Rafetseder (this volume), in the course of their discussion of counterfactual reasoning, they make a series of proposals, some of which are meant to address the point that 3-year-olds' failure on some reasoning tasks involving subjunctive conditionals concerning past events does not seem to arise simply from an inability to inhibit knowledge of what actually happens. First, they suggest that we should distinguish between scenarios, like those employed by Harris, in which (they suppose) the conditional question can be answered correctly on the basis of general background knowledge and/or default assumptions, and those in which it cannot and instead more detailed information about the actual course of events is required. In Harris' muddy shoes scenario, for example, 3-year-olds and other small children will likely have prior experience with the relationship between muddy shoes and floors. According to Perner and Rafetseder, because of this the children don't need to know anything very specific about the actual course of events in Harris' scenario to answer questions about what would have happened if the shoes were removed. They compare the results in Harris' experiments with those from another set of experiments due to Riggs et al (1998). In these experiments, children are told about a firefighter Peter who feels sick and goes to bed. However the fire alarm rings, indicating that there is a fire at the post office and Peter goes off to fight it. When children between the ages of 3 and 4 are asked where Peter would be if there was no fire, far fewer children answer correctly than in Harris' study. Perner and Rafetseder attribute this difference to the fact that in the Riggs et al. experiment, unlike Harris' experiment, providing a correct answer requires referring to the actual course of events in the story—as they put it at one point, if one were to simply ask people, without telling them the details of the story, where Peter would be if the fire had not occurred they would have no idea how to answer, while they would be able to say what would have happened to the floor if

[8] Beck et al. may wish to claim, paralleling their treatment of counterfactual syllogisms, that when reasoning about (5) children employ a fantasy format that reduces inhibitory demands. But for the reason described in footnote 7, this seems ad hoc in the absence of further details.

the doll had walked across it after removing her muddy shoes. They suggest that it is this difference, rather than inhibitory failure per se, which explains why 3-year-olds succeed on some tasks involving conditionals and not others.

Perner and Rafetseder connect this difference to a difference between what they call 'basic conditional' and 'counterfactual' reasoning:

> In both cases, for basic conditional and counterfactual reasoning one has to construct a model in which the antecedent is true and then look at what else is true within that model. This needs background assumptions. In basic conditional reasoning the needed assumptions are made by default, i.e. of whatever is plausible or any suitable assumption that happens to come to mind. In counterfactual reasoning all the details of the actual sequence of events, to which the antecedent is taken to be counterfactual, have to be imported into that model. (p. 100)

Elsewhere they put the distinction in the following way:

> when background assumptions are needed, basic conditional reasoning draws on plausible assumptions while counterfactual reasoning has to take such assumptions from the actual sequence of events (nearest possible world) to which it is supposed to be a counterfactual. (p. 91)

Again I don't see this as establishing a sharp distinction between these two classes of conditionals. To begin with, the examples involving 'basic conditional reasoning' which the authors discuss certainly make use of assumptions of various kinds about what the actual world or the actual course of events is like, whether or not these are explicitly specified. For example, in Harris' example (which Perner and Rafetseder claim involves 'basic conditional reasoning with fictive events'), it is relevant that the actual state of the floor, before the doll walks across it, is that it is clean, it is relevant we assume, either by default or because this is obvious from the way the example is depicted, that the mud is only on the doll's shoes (rather than, say on her feet) and hence will not get on the floor in some other way, even if she removes her shoes, and so on. All of these facts are *specific* facts characterizing the actual situation with respect to which the counterfactual is to be evaluated. It is for just these reasons that philosophers have generally assumed that in evaluating conditionals like those in Harris' example, one must make use of something like the closest possible worlds idea—hence that such conditionals, just as much as the ones that Perner and Rafetseder label counterfactual, involve fitting the assumption specified in the antecedent of the conditional into the course of events and general features that characterize the actual world, with the evaluation of the counterfactual depending on whether its consequent holds under this combination of antecedent and actual world conditions. Even if it is true that some of these assumptions about the actual world are made by default or on the basis of assumed general background knowledge rather than being idiosyncratic to the particular example of interest and explicitly stipulated as such, it is unclear why this should make a difference to whether the conditional in question is treated as a counterfactual or not (or to the kind of reasoning it involves). Moreover, default assumptions often derive from (what are taken to be) typical or common features of the actual world, so

that the whole contrast between what is taken from the actual world and what is assumed by default seems unmotivated. In addition, default assumptions also seem to play a role in the evaluation of the conditionals that Perner and Rafetseder label as counterfactual and not just those they regard as involving basic conditional reasoning. For example, in evaluating counterfactuals about Peter's location if there were no fire, one is likely to assume by default (and even though this is not explicitly specified) that no other fire or event occurs that will cause Peter to leave his bed, and so on.[9]

While Perner and Rafetseder describe the conditionals in the Riggs et al. task in which Peter becomes sick as counterfactuals and those in Harris' experiment as non-counterfactual conditionals and appear to attribute the differential performance of 3-year-olds on the two tasks to this difference in the conditionals involved, it seems to me (at least) equally natural to describe both as engaging in a task involving counterfactual reasoning but which is such that in one case the reasoning involves a familiar subject matter with which the children have considerable experience (muddy shoes and dirty floors) while in the other case, the reasoning involves a less familiar subject matter, with lots of novel detail to keep track of. When the children do better on the first task than on the second, why should we not ascribe their differential performance to these subject-matter-specific differences rather than to some general incapacity to engage in counterfactual reasoning (especially if this requires a redefinition of 'counterfactual reasoning', so that the children in Harris' experiment are not doing counterfactual reasoning)? As we learn from David Sobel's contribution to this volume, there is a great deal of other evidence that (unsurprisingly) children's ability to engage in accurate counterfactual reasoning varies with their causal knowledge concerning the subject matter of that reasoning. Thus children appear to do well with tasks involving counterfactuals concerning simple psychological states such as desire fulfillment and happiness, presumably because they grasp the causal relationship between the two, and much less well with counterfactuals involving other subject matters such as surprise, where they may understand the relevant causal relationships less well.[10]

[9] At the beginning of their paper, Perner and Rafetseder also discuss a number of conditionals associated with scenarios in which there are various rules governing assigning responsibility to the picking of flowers and use this discussion to motivate their claims about the differences between basic conditional and counterfactual reasoning. Space precludes discussion of these examples, but I think their discussion fails to distinguish between questions that are naturally interpreted as having to do with normative rules ('if the gardener does not pick the flowers, the butler should pick them') and conditionals that describe what would in fact happen under various contingencies (e.g. 'if the gardener were not to pick the flowers, the butler would'). That the former questions are answered differently than the latter does not tell us how it is most appropriate to distinguish among different classes of conditionals.

[10] In some respects such an analysis is not terribly far from what Perner and Rafetseder say about the difference between the Harris and Riggs et al. experiments. Perner and Rafetseder suggest that the subjects in Harris' experiments rely more on default assumptions and background knowledge than the subjects in Riggs' experiments. This difference is presumably highly correlated with the extent to which the subjects have prior causal knowledge of and familiarity with the two subject matters. What I find unconvincing is the further inference that the experiments involve different kinds of conditional reasoning, only one of which is genuinely counterfactual.

There is another point that is worth making about the psychological studies under discussion. They rely on a single kind of criterion for whether subjects are able to engage in counterfactual reasoning: whether they correctly answer certain specific verbal questions involving counterfactuals. Even if we confine ourselves to verbal behavior as evidence, one obvious worry about this methodology is that whether subjects are able to correctly answer such questions may be very sensitive to the exact details of their formulation.[11] If this has not already been done, it seems to me that it would be desirable to systematically investigate the extent to which different verbal probes involving counterfactuals or other kinds of conditionals give the same answers. It seems particularly problematic to use the failure of subjects to correctly answer one specific kind of question involving counterfactuals as an indication that they lack general counterfactual reasoning abilities unless there are reasons to suppose that they would also fail to respond correctly to other, related verbal probes. Recall, for example, that in the experiments conducted by Riggs et al. the verbal probe employed to test for the understanding of 'counterfactuals' is 'If he had driven the other way, where would he be?' and 3-year-olds have difficulty with this question. Suppose that the 3-year-olds had instead been asked 'Now imagine that he drove the other way. Where would he be?'. Would the subjects have had similar difficulties with this question? It may be that responses to alternative probes like this have already been explored and children do badly on all of them. However, if it is found that subjects give more accurate answers to some of these other verbal probes than they do to the one employed by Riggs et al., this would seem to me to cast doubt on the claim that failure on the specific verbal probe employed by Riggs et al. indicates a general incapacity for counterfactual reasoning.

A more general issue concerns whether the appropriate tests for whether subjects are able to engage in counterfactual reasoning or thinking should rely only on verbal behavior. One's answer to this question is likely to be influenced by one's views about how broadly one should construe the notion of a 'counterfactual conditional' and 'counterfactual reasoning'. If one uses the notion of 'counterfactual' in the broad way that many philosophers do and thinks of counterfactual reasoning as closely connected with planning, decision-making, and causal reasoning, then it seems hard to deny that non-verbal behavior can sometimes provide evidence for the ability to successfully engage in counterfactual thinking, at least if the behavior in question is sufficiently complex and flexible, under conscious, deliberate control, and is based on 'insight' rather than trial and error learning. An illustration might be provided by an adult who faces a choice among complex plans involving different courses of action, with different outcomes contingent on those courses of action and systematically makes optimal choices on one-shot trials. Here the subject's choice behavior exhibits system-

[11] It is known that a subject's answers to questions involving causal judgments can be strongly influenced by what might seem to be small changes in the wording of verbal probes—see Shanks (2004). A similar point might hold for counterfactual judgments.

atic sensitivity to counterfactual information of the following sort: 'If I were to do A_1, C_1 but not C_2 would result', 'If I were to do A_2, C_3 and C_4 would result', and so on.[12]

Another possible basis for the presence of counterfactual reasoning is neural evidence. For example, Ursu and Carter (2005) find that activity in the orbito-frontal cortex appears to reflect counterfactual comparisons and regret in a choice task, in the sense that this structure is sensitive to comparisons between subjects' actual rewards and alternative rewards that might have been obtained if subjects had chosen differently. Other evidence implicates the dorsal lateral prefrontal cortex in many tasks involving planning and reasoning about alternative courses of action. Particularly if these structures are also active when subjects answer verbal questions about counterfactuals and other kinds of conditionals, this might be taken to show that subjects are engaged in forms of counterfactual reasoning even when they do not answer verbal questions about counterfactuals.

I conclude this section with two additional suggestions about possible research in the general area of counterfactual reasoning. The first, which is implicit in some of my remarks above, is the importance of controlling for potential confounding factors in attempting to discover the explanation for differential performance on various reasoning tasks. When verbal probes are used to assess performance on questions involving counterfactuals (in the narrow sense employed by psychologists), the questions and probes should, at a minimum, be matched for subject matter familiarity, complexity, and so on with other sorts of conditional tasks which are viewed as non-counterfactuals. Some of the questions used in the experiments above meet this standard (such as the pair of questions (7) and (8)), but not all do. If, for example, the children who do poorly on the counterfactual question concerning where Peter will be also do badly on matched future hypothetical questions about a similar scenario, this is an indication that something else besides the counterfactual/future hypothetical contrast explains this bad performance. Relatedly, if young children really suffer from a general limitation in their ability to engage in counterfactual as opposed to future hypothetical reasoning, one should expect this to show up in a variety of different examples and verbal probes.

A second general issue concerns how one might get evidence relevant to assessing the claim that reasoning with future hypotheticals and with counterfactuals (narrow sense) involve different 'types' of reasoning, assuming (as I have argued) that this claim does not follow just from the observation that the conditionals in question differ in various obvious ways. One natural interpretation/operationalization of this claim is

[12] Should all behavior that exhibits, so to speak, sensitivity to alternative possibilities be treated as evidence for counterfactual reasoning? Gopnik (2009) describes babies who recognize that a rake needs to be realigned in order to successfully retrieve a toy as engaged in counterfactual reasoning—they recognize that with the rake in its original position, the toy cannot be retrieved but that if the rake is (were to be) repositioned in just the right way, the toy will (would) be retrieved. I have no strong view about this, but am inclined to think that to the extent that the babies solve this problem in the way that an adult would who consciously and deliberately imagines the rake in alternative positions and then simulates what would follow for retrieval under these alternatives, there is reason to believe the babies (and the adult) are engaged in counterfactual reasoning. The case for this description would be strengthened if the neural areas that are active when an adult solves this task in the way just described are also active in the baby—something that may well not be the case.

that different neural regions or circuits are involved in reasoning with these two sorts of conditionals—a claim that might be assessed by imaging studies in the case of adults or conceivably by lesion studies showing dissociations between the abilities to reason with the two sorts of conditionals. If distinct neural areas or circuits are found, one might then investigate their developmental time course and determine whether it is true that areas associated with future hypothetical reasoning become operational in children before areas associated with 'counterfactual' reasoning—this might then serve as an independent check on the conclusions about this drawn from studies of verbal behavior.

4 Counterfactuals and Causation: Background

The relationship between counterfactuals and causal claims has been a major focus of discussion in both philosophy and psychology for decades. As a point of departure, it will be useful to separate out a number of different more specific issues that fall under this general heading. (I do not claim that these issues are sharply distinct—they tend to blend into each other.)

(1) Issues having to do with providing an account of the semantics (or content or meaning or truth conditions) of causal claims in terms of counterfactuals and/or with whether causal claims imply (in virtue of their meaning) associated counterfactual claims (or vice versa). The counterfactual theories of causation developed in the philosophical literature such as Lewis (1973) and Mackie (1974) have tended to focus on such issues. Lewis' extremely influential account is explicitly reductive: his aim is to provide necessary and sufficient conditions for claims of the form *c* causes *e*, where *c* and *e* are individual events (that is, so-called token causal claims—see Section 6), in terms of counterfactuals involving *c* and *e*, where it (is claimed) these counterfactuals can be understood in a way that does not presuppose causal notions. (In a bit more detail Lewis defines causation in terms of causal dependence: Simplifying slightly, *e* causally depends on *c* if and only if the following two counterfactuals are true: (i) if *c* were to occur, *e* would occur, (ii) if *c* were not to occur, *e* would not occur. *C* causes *e* if and only if there is a chain of events $c, c_1 \ldots c_n, e$ such that each member in the chain (except *c*) is causally dependent on the event before it. Causation is thus the transitive closure of causal dependence which is itself defined in terms of counterfactuals.) It should be borne in mind, however, that even if such reductive projects are unsuccessful (and the relevant counterfactuals require causal notions for their specification), it still might be true that there are interesting and important interrelationships between causal claims and counterfactuals. For example, my own 'interventionist' account of causation, described in more detail below, takes this view—causal claims are connected to a special kind of counterfactual but spelling out the content of these counterfactuals requires (other) causal information. Whether reductive or non-reductive, philosophical accounts that connect causes and counterfactuals have generally focused on what I call below a broad notion of causation—their goal has been to use counterfactuals to distinguish between causal and other sorts of relationships such as those involving mere

correlations. This contrasts with other possible projects, such as employing counterfactuals to characterize ordinary practices of causal selection.

In addition to their possible use to capture features of the meaning or truth conditions of causal claims, there is another, related aspect of the connection between causal claims and counterfactuals that is of psychological interest. This is that, as a matter of empirical psychology, it seems illuminating or clarifying to many people to connect causal claims to counterfactuals. This is reflected in, e.g. the use of frameworks within statistics that explicitly connect causal claims and counterfactuals such as Rubin's potential response model (Rubin, 1974), discussions within social science history that insist that the content of causal claims be made more explicit by connecting them to various counterfactuals they imply, and the use of counterfactuals to clarify the meaning of causal claims in tort law. This connection suggests some interesting questions that are underexplored in both the philosophical and psychological literature: Is the idea that counterfactuals can be used to illuminate causal claims simply an illusion or confusion, as the views of some critics of counterfactual theories of causation might seem to suggest? If not, what does this connection tell us about causal judgment and causal reasoning? Note that on this way of thinking about the relationship between causal claims and counterfactuals, this relationship has a potential *normative* role in the sense that the relationship might be used to clarify, disambiguate, or unpack the content of causal claims (rather than just describing the possibly somewhat unclear ordinary use of such claims). Note also that to accomplish this, we may not need to reduce causal claims to relationships of causal dependence that are non-causal.[13]

(2) Even when construed just as accounts of the semantics of causal claims, counterfactual theories of causation raise issues that are directly relevant to empirical psychology. Most obviously, advocates of counterfactual theories usually claim their theories capture at least many of the causal judgments that people in fact make. Thus it is important to Lewis that his theory reproduce the causal judgments of subjects (or perhaps informed and sophisticated subjects) in cases involving pre-emption, overdetermination, and causation by absence. This raises empirical issues about just what those judgments are. For example, Lewis' theory (at least in its 1973 formulation) implies that in a case in which both c_1 and c_2 are each on their own sufficient to cause e and c_1 and c_2 symmetrically overdetermine e (as when two marksmen simultaneously shoot someone through the heart), then neither c_1 nor c_2 counts as a cause of e. It is far from obvious that this matches the judgments people would actually make about such a case.

[13] I am aware of course that some philosophers will find the idea that counterfactuals might be used to explicate or clarify the content of causal claims as fundamentally misguided since they regard counterfactuals as unclear or metaphysical or perhaps simply think that causal claims are in some way 'prior' to counterfactuals. The alternative view that I am recommending is that it is simply an empirical psychological fact that people sometimes find counterfactuals clearer, more vivid, or easier to understand than the causal claims with which they are associated, or at least that they think they can use the former to clarify the latter. One may then ask what the explanation is for this empirical fact.

(3) A third issue has to do with the psychological processes and representations that underlie human causal reasoning and how these connect (or fail to connect) to the processes and representations that underlie counterfactual thinking. For example, when subjects make causal judgments or internally represent causal claims, do they do this by making or internally representing counterfactual judgments? And when they engage in causal reasoning, do they engage in reasoning about counterfactual claims? Or do causal reasoning and judgment draw on other, quite distinct processes of reasoning and representation that are unrelated to counterfactual reasoning? If, as seems to be the case, the content of causal claims can sometimes be clarified by connecting them to counterfactuals, what if anything does this imply about the psychological relationship between the two?

These issues about psychological processing seem at least somewhat distinct from the issues raised under (1) above because it seems possible that one might be able to provide necessary and sufficient conditions for causal claims in terms of counterfactuals and/or to use counterfactuals to clarify the meaning of causal claims even if, as a matter of empirical psychology, human causal reasoning often proceeds without involving reasoning with counterfactuals. Consider an analogy: A good definition of (or set of necessary and sufficient conditions for) a figure's being a circle is that it is the locus of points that are equidistant from a given point, but, consistently with this, one might think that the definition does not capture how most people conceive of or represent or reason about circles. Perhaps a similar point holds for the relationship between causal claims and counterfactuals. On the other hand, it seems clear enough that many of the philosophers who have defended counterfactual theories of causation, including Mackie and Lewis, have thought that in addition to capturing the semantics of causal claims, causal reasoning at least often involves, as a matter of empirical psychology, reasoning with or about counterfactuals, just as they have supposed that their theories about the semantics of counterfactuals are at least somewhat suggestive about the reasoning that people go through in evaluating counterfactuals.

In addition to these potential connections between causal claims and counterfactuals, there is another observation that will play an important role in my discussion below: human beings seem fairly successful at *learning* about causal relationships. Such success can be manifested in verbal behavior but, as argued in Section 3, it can also be manifested in non-verbal behavior that indicates the acquisition of causal information enabling manipulation and control. For example, a child who correctly selects the object that makes a blicket detector 'go' on the basis of observational evidence involving blickets and non-blickets has learned about a causal relationship, even if he or she cannot correctly answer verbal questions about whether the blicket 'caused' the detector to go off. Psychological theory should explain how such learning occurs (the evidence from which people learn, the learning procedures they employ, the content of what is learned). It is important to recognize that this involves more than merely describing the causal judgments that people make and the factors influencing these. Consider the visual system. It does not just

produce 'visual judgments'; it is also important and in need of explanation that those judgments are often *veridical*. Similarly for causal judgment.

In what follows, I explore some of the issues under (1), (2), and (3) above within the framework of the interventionist approach to causation I have defended elsewhere (Woodward, 2003; hereafter *MTH*). Interventionism was originally intended (at least by me) as a *normative/semantic* account of causal judgment and reasoning—the idea was that one could clarify what causal claims commit us to, the evidence required to support them, and so on by construing such claims as claims about what would happen under possible interventions. However, one can also ask the distinct question of how well this account fares as a psychological theory of human causal judgment. I have begun to explore this elsewhere (Woodward, 2007; see also Bonawitz et al. 2009) and here I continue this project, with special focus on the relationship between causal claims and counterfactuals. (For an overview of interventionism as a psychological theory, see the papers just cited.)

5 Causes and Conditions

When we ask whether a relationship is 'causal', it is important to consider what the quoted word is intended to contrast with—what other kind of relation are we trying to distinguish from those that are causal? One important contrast is with relationships that are merely 'correlational'. The relationship between barometer readings *B* and occurrence/non-occurrence *S* of storms is merely correlational rather than causal; neither *B* nor *S* causes the other; instead both are effects of a common cause (atmospheric pressure). Call a notion of cause which contrasts in this way with 'correlation' a *broad* notion; broad in the sense that it includes relationships involving factors we also readily describe not as causes but as 'conditions', or as 'background' or 'enabling' factors for an effect. The relationship between the presence of oxygen and the outbreak of fires is 'causal' according to this broad notion, even though we also describe oxygen as a condition rather than a cause of the fire. To mark this difference, let us call a notion of cause that contrasts with conditions, enabling factors, etc., a *narrow* conception of 'cause'.

We may distinguish two extreme positions regarding the narrow cause/condition contrast. One possibility is that this contrast is entirely 'objective', depending just on features 'out there' in the world, rather than on 'subjective' considerations such as the inquirer's attitudes or interests. A simple example of an objectivist view is the claim that conditions are causes in the broad sense that (in addition) are nearly always present (in the contexts of interest), such as the presence of oxygen in the case of fires. Conditions in this sense might then be contrasted with causes in the narrow sense (narrow causes) where these are understood as factors that are causes in the broad sense that are sometimes present and sometimes not, such as strikings of matches. This is not a very adequate account of the cause/condition contrast but it does ground the contrast in features in the world.[14]

[14] For example, it fails to account for the experimental results from Mandel discussed below.

An alternative possibility is that the narrow cause/condition contrast has to do entirely with features of the mental states of inquirers—e.g. with what they find interesting or salient: The policeman regards the drunkenness of the driver as a cause of the accident and the inadequately banked curve as a condition because of his professional role/interests, while the highway engineer regards the inadequately banked curve as among the causes because of *his* interests, but there are no features of the world that ground these differences in judgment. Of course another possibility is that some mixture of these extreme positions is correct.

What would follow for the empirical psychology of causal cognition if the cause/condition contrast turned out to have little or no 'objective' basis? This depends on what we are interested in explaining. Presumably, it would still be of psychological interest to ask about the subjective factors that influence how the cause/condition distinction is drawn and how this distinction influences other aspects of cognition, such as moral judgment. On the other hand, if the distinction is entirely 'subjective', focusing on it may tell us little about how we are able to successfully learn about relationships that are useful for purposes of prediction and control, (assuming, as seems plausible, that such relationships have to do, at least in part, with features of the world that hold independently of people's interests). Thus for the purposes of understanding learning, there may be much to be said for focusing, at least in part, on a broad notion of cause, since the cause/correlation distinction is arguably both 'objective' and central to our interest in manipulation and control.

The process of distinguishing, among those factors which are broad causes, those that we regard as causes in the narrow sense and those that are mere conditions is often described as *causal selection* and is the focus of a great deal of research, particularly in social psychology. I do not dispute that this is a worthy project for psychological inquiry, but I do claim it is *different* from elucidating the broad cause/correlation contrast, and that an account of causation may be illuminating in connection with one of these projects and not the other. In particular, a proposed connection between causation and counterfactuals may be helpful in distinguishing between broad causal relationships and correlations, but not in understanding causal selection. This is how Lewis conceives of the counterfactual theory presented in his paper (1973): it is explicitly described by Lewis as an attempt to capture a broad notion of causation having to do with particular events, and *not* as a theory of causal selection. This is also true of the account presented in *MTH* and the condition **(M)** connecting causation and interventionist counterfactuals described in Section 7.

6 Actual Cause Judgments

The philosophical literature distinguishes between *type causation* (or type causal claims) and *token causation* (or token or singular causal claims). Type causal claims are often thought of as general causal claims making no explicit reference to any particular instantiation of the cause and effect, as when one says that short circuits cause fires,

without referring to any particular short circuit or fire. Token causal claims, by contrast, explicitly claim that some particular event was causally responsible for another, as when one says that the short circuit occurring in Jones' attic at time *t* caused the fire that destroyed the attic at *t+d*. The type/token terminology is deeply entrenched in philosophical usage. Nonetheless, as ordinarily understood, it is misleading (cf. Woodward, 1990, Hitchcock, 2007). First, contrary to what the type/token terminology suggests, type causal claims should *not* be understood as claims about the existence of causal relationships between abstract types or properties, which are not realized in particular occurrences. When short circuits cause fires, this always involves particular individual short circuits (tokens) causing particular individual fires (also tokens).

A deeper point concerns the relationship between type causal claims and our practices of ascribing responsibility to particular events in causing others, which in order to avoid some of the unclarities of the type/token distinction, I will call *actual cause judgments*. As many examples in the literature show, one may have complete knowledge of the type causal relationships obtaining in a given situation (telling us how *token* realizations of those types are causally related in the situation) and yet this information does not seem to fully constrain the actual cause judgments we make. Suppose that smoking causes lung cancer (a type causal relationship), Jones smokes heavily, and Jones develops lung cancer. These assumptions leave it open whether Jones' smoking was an actual cause of his lung cancer—this might be true, but it might also be true that the actual cause of Jones' lung cancer was something completely different such as his exposure to asbestos. For similar reasons, adding that Jones was exposed to asbestos and that such exposure causes lung cancer also does not settle what caused Jones' lung cancer. More complex and subtle examples illustrating the same point can be found in Collins, Hall, & Paul (2004).

Although I lack the space for detailed discussion, I take these considerations to suggest that type cause and actual cause judgments play at least somewhat distinct roles in our thinking and often may be guided by different sorts of considerations. With type causal judgments our concern is often forward-looking: it involves reasoning *from* a cause or potential cause *to* its effects. Often this involves a concern with prediction or manipulation or with explaining repeatable events. For example, we may use type causal information to predict what will happen if a certain kind of cause occurs or to tell us which means to introduce in order to bring about some desired outcome, as when we reason that application of fertilizer will cause crops to grow or putting a certain object on a blicket detector will activate the detector.

By contrast, actual cause judgments are often made in cases in which a particular outcome occurs (often but not always, something bad). Our concern is then to reason 'back' *from* this outcome *to* another occurrence which can be regarded (either literally or figuratively) as 'responsible for' or 'to blame' for it. For example, we see a stunted plant and ask what is responsible for its deficient growth (as opposed to asking whether application of fertilizer will in general stimulate plant growth). In some cases, the responsibility in question is moral or legal; in other cases, responsibility may be assigned

on the basis of expectations about normal biological functioning or on ideas about what counts as a deviation or disruption of such functioning (e.g. disease) or on the basis of some engineering or design norm, as when we single out the behavior of the O-rings as the cause of the Challenger disaster.

I don't claim that this conceptualization in terms of responsibility captures everything that underlies our practices of actual cause judgment, but merely that this is one factor that seems to shape these judgments. However, if this is correct, it suggests it may be a mistake to suppose that getting clearer about our practices of actual cause judgment will automatically illuminate every aspect of our practices involving type cause judgments. In particular, it may be a mistake to suppose that type causal judgments are merely generalizations of types of actual cause judgments; instead the two kinds of judgment, although not completely unrelated, may be guided by considerations that are at least somewhat distinct.

Why does this matter? If one looks at the psychological literature on causal cognition, one sees what looks like two rather different research paradigms or projects, paralleling the distinctions drawn above. In one, typified by the research reported in Gopnik and Schulz (2007), the focus is mainly on learning and reasoning about type causal relations and the causal relations of interest are 'causal' in what I called the broad sense. Normative theories of causal learning and reasoning are relevant in this paradigm because *successful* performance is part of what researchers want to understand. In a typical experiment, subjects are put in an environment in which there are unknown causal relationships and where there is a well-defined notion of making a mistake about these—e.g. concluding something makes a blicket machine go when it doesn't. Successful acquisition of causal information may be shown by non-verbal behavior (picking out one object rather than some alternative as the one that will activate a detector) as well as by the correct use of words like 'cause'. Because of this, the experimental paradigms allow for the possibility (at least in principle) that subjects may have causal representations (and may support these with processing of counterfactuals, probabilities, etc.) even if they are non-verbal.

In another, contrasting paradigm (common in social psychology), experiment and theory focus largely on understanding the criteria governing actual cause judgments (and in particular understanding causal selection in connection with such judgments). In a typical experiment (see, e.g. the example from Mandel below) subjects are presented with verbal scenarios involving a particular outcome in which all the broad causal relationships (involving many different factors) are either stipulated or in some cases left unspecified. The researcher's interest is which of these factors subjects will pick out as 'the cause' or will rate as most strongly causal and what factors affect such judgments. There is no other basis for identifying subjects' actual cause judgments besides their verbal responses to questions. Subjects do not face a learning problem in the way that subjects in the first paradigm do, and there is no obvious standard for whether the subjects' judgments are correct other than whether they conform to the most common patterns of attribution.

I repeat that in my view both projects are interesting and worthwhile. However, I also want to stress that that the understanding of actual cause judgment (including how these are connected to counterfactuals) that the second paradigm yields may tell us less than we would like to know about the project of understanding how humans acquire and use type level causal knowledge for manipulation, control, and explanation, and how type causal judgments are connected to counterfactuals.

As an illustration, consider an interesting paper by Mandel (2003). Mandel is interested in the role that 'counterfactual thinking' plays in causal selection, particularly as expressed in actual cause judgments. He considers the suggestion that those factors most readily picked out as the (or a) cause of some outcome are just those factors that figure in the antecedents of those counterfactuals that most readily come to mind when subjects are asked to consider counterfactuals about which factors might be changed or mutated to prevent or undo the outcome.

Mandel has little difficulty in showing that (at least when taken in an unqualified way) this suggestion is mistaken. In one experiment (originally reported in Mandel and Lehman, 1996), subjects are given a scenario in which Jones decides to drive home via a new route and is injured in a collision with a drunk driver. When subjects are prompted to consider counterfactuals leading to the undoing of the accident, they focus most readily on counterfactuals in which Jones does not choose the new route, but when asked for the cause of Jones' injuries, they select the collision. This suggests that people's causal selections are not guided (or are not guided only) by a particular kind of 'but for' counterfactual, one that involves undoing the effect. In this sense, there is, as Mandel puts it, a 'dissociation' between causal and counterfactual judgments.

Does it follow that counterfactual theories of causation are mistaken *in general* when taken as psychological theories? This conclusion seems premature. First, Mandel's results have to do with one particular aspect of causal cognition: causal selection in connection with actual cause judgments; they simply don't address the possible role of counterfactuals in discriminating between broad causes and correlations in type cause judgments. Moreover, in the latter connection, what matters, according to most counterfactual theories of causation, is which counterfactuals are *true* (or are regarded by subjects as true), not which counterfactuals come most readily to mind. It would be very surprising if Mandel's subjects were to deny the counterfactual claim that if the collision with the drunk driver had not occurred, Jones would not have been injured —subjects most likely think that this counterfactual is true, even if it is not the counterfactual that most readily comes to mind when people are asked to 'undo' the accident. In this respect there *is* a connection (rather than a 'dissociation') between the causal role of the drunk driver and a counterfactual linking the driver and the injury. Mandel's results are thus consistent with the contention that counterfactual thinking figures in the distinction people draw between those factors that are broadly causal (these will include both the collision and the choice of route) and those that are merely correlational. Nor, for reasons already explained, do the results of such experiments tell us much about how we learn (or reason) about type causal relationships.

Subjects are *told* what all of the relevant broad causal relationships are in the scenario (there is no learning), and it is not obvious in what sense someone who selects Jones' choice of route as a cause is making a mistake, other than in departing from common usage.

Another point concerns what it means to hold a 'counterfactual' theory of causation. There is a strong tendency in the psychological literature (illustrated by Mandel's scenario) to focus on hypotheses about the connection between causal claims and a particular kind of counterfactual: a 'but for' or 'necessary condition' counterfactual that relates the absence of the cause to the absence of the effect. (This is understandable since in both philosophy and the law counterfactual theories of causation have tended to employ such counterfactuals). It is thus important to realize that there are many other varieties of counterfactuals that may be linked to causal judgment in one way or another. For example, it is certainly possible to use counterfactuals to attempt to capture the idea that causes are *sufficient* (as opposed to necessary) in the circumstances for their effects:

If *C* causes *E*, then if *C* were absent from the actual circumstances and then introduced (via an intervention—see below) into those circumstances, *E* would occur.

To avoid trivialization, we would also need to provide truth conditions for this counterfactual according to which it is not automatically true as long as instances of *C* and *E* occur (i.e. we need to reject a 'strong centering' principle for counterfactuals, but, as noted above, there are independent reasons for doing this—see footnote 3.) Other causal notions and distinctions can also be captured counterfactually, again using counterfactuals that are not 'but for' counterfactuals (Woodward, 2006). The larger point is that although it is true that 'but for' counterfactuals do not capture all aspects of causal judgment, it does not follow that some other features of causal judgment cannot be captured via other counterfactuals.

7 Interventionism

With this as background, let us turn to some issues in the psychology of causal cognition specifically raised by interventionist accounts of causation. Here I will provide only the briefest overview—the reader is referred to Woodward (2003, 2007), as well as the other papers in Gopnik & Schulz (2007) for additional discussion. Interventionists think of causal relationships as relationships that are potentially exploitable for purposes of manipulation and control. A simple characterization of what it is for *C* to be a type cause of *E* within such a theory is:

(**M**) If an intervention that changes the value of *C* were to occur and if under such an intervention (and no others) *C* and *E* would be correlated (or equivalently if *E* would change under this intervention on *C*), then *C* causes *E*. Conversely, if *C* causes *E, then* there are possible interventions that change the value of *C* such that if such interventions (and no other interventions) were to occur, *C* and *E* would be correlated (equivalently, *E* would change under this intervention on *C*.)

'Intervention' is a technical notion; it is a generalization of the notion of an ideal experimental manipulation. Manipulations carried out by human beings are among paradigmatic cases of interventions (and such manipulations arguably play an important role in causal learning—see below) but the notion of an intervention itself can be specified without reference to human agency. However, causal notions *are* required for the specification of an appropriately behaved notion of intervention. Thus while **(M)** yields a link between causal claims and what happens under interventions, it does *not* yield a reductive analysis of causation in terms of claims about counterfactuals (or anything else.) To the extent that we are willing to use (**M**) as a basis for a psychological theory, it should *not* encourage us to think that it is possible to characterize causal beliefs in terms of a more general category of counterfactual beliefs that are non-causal in character.

(**M**) is specified in terms of a particular kind of counterfactual which I will call an *interventionist counterfactual* having to do with what would happen to *E* if interventions on *C* were to occur. (The notion of a counterfactual at work here is the broad, philosophical notion described in Section 2, rather than some narrower notion.) This counterfactual formulation seems natural since we do not want to make whether *C* causes *E* depend on whether an intervention on *C* actually occurs.

(**M**) is intended as a characterization of a broad notion of (type) causation; it plainly fails to capture the narrow cause/conditions contrast. (For example, it counts both the occurrence of short circuits and the presence of oxygen as among the causes of fires.) For reasons explained above, I do not regard this as a fatal deficiency in (**M**), either when construed as a normative or descriptive psychological theory. (**M**) will be defensible in the latter role, as long as people draw a contrast between causal and correlational relationships and **(M)** correctly captures this.

How should interventionist counterfactuals be understood and what, so to speak, is their psychological significance? As explained above, in normative versions of the theory, the notion of an intervention is not linked essentially to human action. Nonetheless, in developing the approach into a psychological theory, it is very natural to assign special significance to those interventionist counterfactuals that are connected to action (and to planning and means/ends reasoning) directed at changing or transforming the world in order to achieve some goal. If interventionism (construed as a psychological theory) is on the right track, grasp of such conditionals will be central to at least some aspects of causal cognition.

In pursuing this idea, two crucial questions are (i) whether it is possible to be 'implicitly' guided by or sensitive to the information provided by such action-oriented counterfactuals even if one is not able to provide correct verbal answers to explicit questions about them, and (ii) supposing that this is possible, what would constitute evidence for such guidance. The motivation underlying these questions is that, as suggested above, it seems natural to understand a great deal of planning and goal-directed action (including that exhibited by small children) as involving a (perhaps tacit or implicit) appreciation of counterfactual (or at least conditional) relationships. Consider an agent *A* who recognizes that the present state of the world is *X* and who wishes

to achieve some goal which involves the world being in state $X' \neq X$. Suppose also A recognizes that by doing Y, she can change the state of the world to X'. Here we have a simple, action-directed counterfactual (that is, a counterfactual in the broad philosophical sense—call it a future hypothetical if you would prefer): if A were to do Y, then X' would be the case. Although this assumption may prove, on further analysis, to be mistaken, it seems very natural to suppose that in some circumstances A's non-verbal behavior can provide evidence that she is guided by or sensitive to the information in such conditionals/counterfactuals and that this can occur in subjects, such as young children, that lack the cognitive and linguistic sophistication to accurately answer many verbal questions about such conditionals. In addition to the example from Gopnik cited in Section 3, a number of other phenomena, including backward blocking and learning involving combinations of interventions seem to suggest a similar interpretation in terms of implicit processing involving conditionals.[15]

Let us tentatively adopt such an interpretation and see where it leads. First, note that a similar sensitivity to action-counterfactuals must also be present among subjects who are able to parse or minimally understand the actions of others, at least in so far as such parsing involves decomposing other's actions into goals or ends, on the one hand, and the means employed to achieve these, on the other. For example, Meltzoff (2007) has shown that even very young infants can distinguish between an adult model's goals (pulling an object apart, activating a detector) and the means the model employs in attempting to achieve these, adopting different means when it is appropriate to do so. Such action parsing and means/ends discrimination is known to emerge very early (Meltzoff, 2007) and a natural thought is that it might be part of a scaffolding enabling acquisition of a full-fledged interventionist conception of causation and the understanding of conditional relationships that accompanies it. This idea is explored in more detail in a number of the papers in Gopnik & Schulz (2007).[16]

[15] For such an interpretation of backwards blocking, see Mitchell et al. (2005) and for additional discussion, McCormack et al. (2009).

[16] As an illustration of how children's actions in response to an adult model naturally suggests a description in terms of counterfactual processing, consider the following account from Perner of an experiment of Meltzoff's:

When observing an adult model trigger the blinking of a box by touching the box with her forehead (Meltzoff, 1988) infants will also use their head (and not their hands) for that purpose, but only if the model *could have used* her hands. Infants do use their hands when the model's hands were occupied otherwise (for holding a blanket) during the demonstration. Underlying this difference between conditions must be infants' counterfactual conditional reasoning: 'if the model *had had* her hands free'. (Perner, Sprung, & Steinkogler, 2004: 197f)

Since the infants are *pre-verbal,* Perner's idea must be that the infant's conditional reasoning is in some sense 'implicit' or at least that its ascription does not require that the infant be able to correctly answer questions posed in the form of conditionals (Is it true that if the model could have used her hands, then . . . ?). To anticipate an objection, I agree we should distinguish cases in which one is genuinely guided by an implicit grasp of a counterfactual, from cases in which the subject has merely acquired an action-response pattern as a result of instrumental conditioning. But since Perner's example involves one-shot imitation learning, it does not seem plausible to think of it just as a case of instrumental learning.

A second set of issues concerns the character of the conditionals associated with **(M)** and with goal-directed action. I noted above that these will count as counterfactuals according to the standard philosophical conception of counterfactuals. However if one adopts the narrower usage favored by psychologists, then these conditionals seem to be mainly or entirely future hypotheticals—that is, claims about what will happen, under the assumption that certain interventions occur in the future. Thus if beliefs about type causal relationships require or in some way involve a grasp, implicit or otherwise, of conditionals (and one wants to distinguish genuine counterfactuals from future hypotheticals), a grasp of future hypotheticals may be sufficient.

This is relevant to psychological issues about the relationship between causal beliefs and mastery of various kinds of conditionals among young children. Suppose, for the sake of argument, that we accept that (a) young children have difficulty understanding and correctly evaluating the sorts of conditionals that psychologists describe as counterfactuals or at least that they find these considerably more difficult than the corresponding future hypotheticals. Nonetheless (b) such children exhibit an understanding of a range of causal concepts—as Roessler (this volume), observes, they understand concepts like push, pull, etc. and use them to make true causal claims, and they correctly solve various other kinds of causal inference problems. It may seem tempting to infer that at least for such children causal beliefs and understanding do not consist in the mastery of counterfactuals—that there is a 'dissociation' between children's abilities to make causal judgments and their capacity for counterfactual reasoning.[17]

The observation that the representation of type causation in terms of conditionals may require only the representation of future hypotheticals may go some way toward reconciling (a) and (b). That is, if the sorts of conditionals associated with the understanding of type causal claims are future hypotheticals, then to the extent that young children do exhibit mastery of these, there may be no 'dissociation' between their understanding of conditionals and their understanding of causal concepts.

Interestingly, though, matters may be different with actual cause judgments. Here, precisely because such judgments commonly require identifying a cause of an event that has already occurred, to the extent that an account of such judgments in terms of conditionals is possible at all, it seems plausible that these will be 'counterfactuals' in the (narrow) sense favored by psychologists—that is, they will be conditionals about the past whose antecedents involve the supposition that some actually occurring event did not occur. For example, if we want an account that connects the actual cause judgment that the short circuit caused the fire to a conditional of some kind, this will presumably be (or at least include) a conditional concerning what would have happened if the short

[17] For discussion of some of the relevant empirical results, see McCormack et al. (2009). Unlike some other researchers, these authors find evidence that counterfactual and causal judgments are 'psychologically related', although the former lag behind the latter in development. In their words, their findings 'suggest that whatever account is provided of the processes and/or representations underpinning children's causal learning on the blicket detector task must be one that can explain children's ability to make not only causal judgments but also counterfactual ones' (McCormack et al., 2009: 1573).

circuit had not occurred. Thus to the extent that reasoning (whether implicit or explicit) with such counterfactuals is required for actual cause judgments, and to the extent that such reasoning is more difficult for young children than reasoning with future hypotheticals, then it is arguable that we should expect young children to have more difficulty with actual cause judgments than with type cause judgments and to acquire the ability to use the latter appropriately before they acquire the ability to use the former. To the best of my knowledge this issue has not been explored empirically—it would obviously be desirable to do so. If my remarks on the differences between actual cause and type causal judgment are on the right track and if, for whatever reasons, future hypotheticals are easier for children than counterfactuals with false antecedents concerning the past, it seems entirely possible that there may be some developmental dissociation between these two kinds of causal judgment, with mastery of the latter preceding mastery of the former. Of course, another possibility is that a capacity to reason with counterfactual conditionals (in either the sense favored by psychologists or the sense favored by philosophers) is not required for actual cause judgments and that some other form of reasoning is involved in such judgments. And yet another possibility is that once one controls for subject matter and other factors, children do just as well with future hypotheticals as with counterfactuals in the narrow sense, in which case they may do just as well with actual cause judgments as with type cause judgments.

8 Implicit Knowledge of Counterfactuals Again

Putting aside issues about differences in the ways in which type and actual cause judgments may be related to conditionals of various sorts, let me return to the more general question of how explicit, conscious, and/or tied to verbal performance a subject's cognitive processing needs to be to qualify as 'counterfactual reasoning' (in either the psychologist's or the philosopher's sense). When young children fail to give correct answers to questions involving counterfactuals, does this by itself show that they are unable to engage in counterfactual reasoning or that it is wrong to think their causal beliefs involve implicit grasp of counterfactuals? Or can counterfactual reasoning/processing be manifested in non-verbal behavior?

In considering this issue, it is worth observing that a similar dissociation between apparent implicit processing and explicit verbal performance is common elsewhere. There are a number of tasks in which both infants and adults process statistical information and use it to modify their behavior and judgment in normatively correct ways (e.g. Saffran et al. 1996). At the same time, even adults are notoriously bad at many problems involving probability and statistics when these require conscious deliberation and explicit verbal answers. The general pattern seems to be that often abilities involved in implicit processing are not easily deployed in explicit reasoning tasks—good normative performance on the latter requires specialized forms of explicit training, such as a course in statistics. A natural thought is that a similar point might hold for reasoning involving conditionals, including counterfactuals.

Of course, the notion of implicit representation/processing in general (quite independently of whether it is claimed to involve counterfactuals) is far from transparent. This is an old issue in cognitive science and I won't try to resolve it here. I will observe, however, that virtually all major theories of causal learning and judgment in psychology are not even prima facie plausible if they are understood as claims about a subject's explicit mastery of concepts occurring in those theories. For example, the claims that a subject's causal judgments track $\Delta p = P(O/A) - P(O/\text{-}A)$ or that they track causal power in Patricia Cheng's (1997) sense are not remotely plausible if understood as claims about a subject's abilities to answer explicit questions about conditional probabilities. It would be absurd to claim that Cheng's theory is refuted just on the grounds that typical subjects are unable to understand the expression for the causal power of a generative cause ($=\Delta p \,/\, 1\text{-}p(e/\text{-}c)$) occurring in her theory or make mistakes when they attempt to evaluate this expression. In this respect, the claim that subjects engage in implicit counterfactual processing when they make causal judgments seems no different from similar claims made by other theories of causal judgment. It is a claim about the existence of certain patterns or regularities in their reasoning, behavior, and responsiveness to evidence, and not a claim about the subject's introspective access to or understanding of those patterns (or about the subject's introspective access to descriptions of those patterns provided by theorists).[18]

It is also worth noting that these considerations bear on the suggestion that, assuming that it is correct that counterfactual reasoning is difficult for small children, this is because of the demands it imposes on processes of working memory and executive function that are underdeveloped in small children. It may be that such difficulties in executive functioning and working memory are more likely for tasks involving explicit counterfactual reasoning (requiring correct verbal responses) than for tasks involving implicit counterfactual processing, assuming we can make sense of the latter idea.

9 An Interventionist Account of Causal Representation

My argument so far has been, in effect, that any attempt to connect causal cognition with counterfactual processing is going to have to regard much of that processing as 'implicit' if it is going to be remotely plausible. That said, it seems undeniable that, even from an interventionist perspective, more is involved in human causal cognition than

[18] This way of looking at things may help with another puzzle. On the one hand, it seems introspectively plausible that running what one might think of as counterfactual simulations of what would happen if ... in one's head can sometimes help one to reason about causal relationships. On the other hand, in 'running' such counterfactuals and seeing what results, we seem to make use of causal knowledge that we already possess, at least on an interventionist treatment of causation. If so, how could this process possibly help to clarify our understanding of causal relationships? One possible answer: The use of counterfactual simulation might help to clarify causal relationships if the causal/counterfactual knowledge one draws on to run the simulation is tacit or non-explicit, and if the simulation allows one to access or make explicit implications of this knowledge. A similar idea might help to explain why (as a matter of empirical fact) connecting causal claims to explicit counterfactuals often seems to help clarify or elucidate them.

just being guided in one's behavior by implicit processing involving counterfactuals linking one's own actions to outcomes. For one thing, the above description may be satisfied by learning involving just operant conditioning—that is, the implicit learning of action/outcome contingencies, and this seems to fall short of full-fledged human causal representation.

What more, then, is required for guidance by genuinely causal representations? Let me suggest two candidates for additional *necessary* conditions (I do not claim that they are jointly sufficient).

(1) To the extent that operant conditioning just involves an organism learning about relationships between its own behavior and outcomes these produce, the representations involved are egocentric. Adult human causal representation is *allocentric* rather than egocentric; it involves the idea that the very same relationship present between my actions and outcomes can be present both when other organisms act, and in nature, independently of the action of any organism. To employ an example from Woodward (2007), human causal representation is such that it represents the same (type) causal relationship as present between (i) the shaking of a tree limb by the wind and the fall of fruit from the limb and (ii) when we or another organism shake the limb. This helps to enable forms of causal learning in humans that would not be possible in organisms whose representations were more egocentric—one can learn both from the interventions of others and from observing regularities in nature.

(2) Human causal representation seems map-like in other respects as well. Although type causation is not in general transitive, it is often true that if (i) X causes Y and (ii) Y causes Z, we may reliably conclude that (iii) X causes Z. Human planning and means/ends reasoning to produce desired effects often involves (among other things) the capacity to put together beliefs of form (i) and (ii) together to get a belief and associated procedure in the form of (iii) which allows one to produce some desired goal: I want my enemy dead and recognize that a large falling rock will kill him but I can't directly move the necessary rock. I also recognize, however, that levers can move rocks, etc. Notice that this causal connection between moving the lever and my enemy's death may not reflect any regularity that I have previously observed occurring spontaneously in nature. Instead, I likely have to (physically) construct the relevant connection after (in some sense) recognizing its possibility. This is analogous to the way in which possession of an allocentric spatial map can allow for the solution of novel navigational tasks that could not be solved with a more egocentric form of spatial representation. Like allocentric spatial maps, human causal representation takes a form which integrates or ties together different pieces of causal information rather than leaving them isolated from one another and thus allows causal insight into novel situations.

10 Causal Perception

Among the many issues that I have neglected so far is the role of so-called causal perception in establishing causal conclusions. This is particularly emphasized in Johannes

Roessler's contribution to this volume. Roessler suggests that many of the causal judgments we make in ordinary life are established on the basis of perceptual clues and, following Anscombe and Strawson, advocates a particularist conception of causal understanding based on 'thick' specific causal concepts.

It is uncontroversial that there is a range of 'mechanical' interactions in which adult subjects report perceiving that a causal interaction has occurred—these include Michottean collisions or 'launching events' in which one object strikes another and causes it to move, phenomena in which there is a perception of pulling, pushing, or bursting, and arguably perceptions of support. At the same time, there are many other cases in which subjects make causal judgments and in which causal perception and the spatio-temporal/mechanical cues associated with it are not sufficient to disambiguate competing causal hypotheses. This can readily happen even with simple mechanical devices—for example, in the case of the gear toy discussed in Gopnik & Schulz (2007) spatio-temporal/mechanical cues are insufficient to tell subjects whether one gear causes the other to move or both are caused to move by a common cause. Here active intervention (which is successfully performed by small children) is required to sort out the causal structure of the device. In other situations, including many encountered in ordinary life, while the interactions involved may be mechanical at some microscopic level, causal perception of the sort under discussion is not helpful in discovering what the causal relationships are. This is generally true, for example, if one is trying to determine whether certain agricultural techniques enhance crop growth, whether ingesting a food will make you nauseous, whether exposure to a putative pathogen will make you ill, or whether a structure is sturdy enough to hold a certain load or withstand certain conditions of use and weather.[19] In these cases, observation of covariation between candidate causes and putative effects—whether this involves passive observation or observing the results of one's own interventions (or those of others)—plays an important role. In addition, a huge body of evidence shows that human beings (including small children) can learn causal relationships, including arbitrary and unfamiliar causal relationships that are not subsumable into previously familiar categories of pushing, squashing, and so on, from covariational information. For this reason, while I agree that causal perception plays a role in ordinary causal learning and judgment, my view is that it is simply not true that all or most causal learning or judgment in ordinary life just involves the exercise of causal perception or the application of such already learned 'particular causal concepts'. In addition, and

[19] The border between perception and cognition is notoriously slippery and this may make it tempting to expand the notion of causal perception to include cases of what many would regard as non-perception causal based judgment (e.g. judging that a drug cures a disease). Against this tendency, I would argue that the neural structures and processes underlying uncontroversial cases of causal judgment are largely distinct from those underlying non-perceptual causal judgment. Associated with the former (in cases of the visual perception of causation) are areas like MT and STS, which we have independent grounds for regarding as visual or sensory processing areas. Causal judgment that is not based on causal perception involves more frontal areas like DLPFC, which insofar as there is a cognition/perception distinction at all, falls on the cognitive part of this divide. It is not true that all causal judgment activates areas involved in causal perception narrowly conceived.

quite independently of what happens most often,[20] it is a crucial feature of human causal learning that we are able to learn novel causal relationships that do not fit into pre-existing specific categories. Thus if the particularist account of causal understanding advocated by writers like Strawson and Anscombe amounts to the claim that any general concept of causation is entirely parasitic on more causal specific concepts, that the more general concept plays no important independent role in ordinary human cognition, and that causal learning from covariational information is unimportant, I think that such an account is mistaken on empirical grounds. Relatedly, the particularist account seems to me to completely fail to explain causal learning, both because it has nothing to say about how novel causal relationships are learned and because, even when causal judgment takes the form of applying pre-existing particularist categories to particular situations (so that various episodes are classified as scrapings, squashings, etc), the particularist account says nothing either about how these categories are initially acquired or how we learn to recognize when particular episodes fall under them.[21]

I can further clarify these points by commenting briefly on some remarks in an earlier draft of Roessler's chapter in this volume,[22] in which he comments on a passage from my *Making Things Happen*:

> Here is Jim Woodward, commenting on the example of 'an early hominid who observes that rocks tend to move when struck by other rocks of roughly equal size in cases in which such impacts occur "naturally", without human intervention': 'the possibility that the relationship in question [is] merely correlational seems so unlikely that it is easy to overlook the point that a separate inferential step is required to move from the claim that two kinds of events are correlated to a claim about how one will respond under manipulation of the other.' Suppose we insist that characterizing the early hominid's experience as a sequence of experiences of

[20] Even if it were true that most causal judgment in everyday life takes the form of the application of previously learned thick causal concepts on the basis of causal perception, it does not in my view follow that theories of causal learning and judgment should focus exclusively or primarily on such episodes. The episodes that occur most frequently in ordinary experience are often not those on which it is most worthwhile or illuminating to focus in constructing a scientific theory. Ordinary everyday occurrences of motion are very complicated—one learns more by focusing on statistically rare or abnormal cases like moving pendulums, or objects falling in a vacuum. Similarly for causal judgment/learning, with the added point that to the extent that we just focus on the application of already learned thick causal concepts and relations, we ignore the important problem of how these were learned in the first place.

[21] For additional discussion of the role of causal perception in causal learning and judgment, see Woodward (2011). It is worth remarking that when adults make causal judgments on the basis of perceptual cues, they seem to employ a concept or representation of causation that integrates perceptual elements with elements having to do with difference-making or counterfactual dependency: thus in judging on the basis of perceptual cues that a collision with one billiard ball caused a second to move, they typically also judge (assuming that the situation is one in which other causes of motion are absent) that if the collision had not occurred, the second would not have moved. For this reason, it seems to me to be a mistake to suppose that causal judgment based on perceptual cues is something entirely distinct from the sorts of judgments that are associated with a dependency-based conception of causation.

[22] This passage does not occur in the final version of Roessler's chapter, which was made available to me only after completion of my draft. A related discussion occurs on p. 82f of his chapter, but this raises a number of issues that are tangential to the points made above.

> distinct events—an experience of a stone striking another, followed by an experience of the second stone moving—is phenomenologically off-key: if the hominid is at all like us, he will experience the event as a specific mode of causal transaction, as a single event or process of one stone knocking off another. So the question of whether the relationship between two observed events is causal or merely correlational simply will not arise.

First, in claiming that a separate inferential step is required to move from correlational information to causal conclusions, I was not making a claim about anyone's phenomenology. My point was rather a matter of logic: In many cases, the same correlational evidence is consistent with a number of different causal claims—hence additional information or assumptions are required to get from the correlational information to a unique causal conclusion. For example, the existence of a correlation between *X* and *Y* is consistent with *X* causing *Y*, *Y* causing *X*, or *X* and *Y* having a common cause, among other possibilities. This does not mean or require that when someone judges in this case that e.g. *X* causes *Y*, they must as a matter of conscious phenomenology first go through a 'precausal' stage of awareness in which *X* and *Y* are represented as merely correlated and then on the basis of what is represented at this stage, consciously go through a series of separate inferential steps which lead to the conclusion that *X* causes *Y*. Sometimes this may be a reasonable description of what happens but in other cases, just as Roessler claims, the subject may not be consciously aware of going through the processing just described. This is true of the cases we think of as involving causal perception: In such cases it seems to the subject as a matter of phenomenology that she directly sees that *X* causes *Y* (or that some particular instance of *X* causes *Y*), with no awareness of a distinct presentation of evidence involving correlations and no awareness of intermediate inferential steps. However, this does not mean that prior processing fails to occur or that it is not the business of psychology to describe it. Even in the case of launching phenomena, which presumably represent the kind of case that is most favorable to Roessler's point of view, there must be some computational story about how the visual system, in response to various spatio-temporal cues (and perhaps covariational information as well), produces the perception that some interactions are causal and others are not. Subjects may not be consciously aware of such cues or may not be aware of them as distinct spatio-temporal cues that occur independently of the perception of causation, and certainly will not be aware of the computations that generate this perception of causation, but this does not mean that it is misguided to appeal to these factors to explain how this perception occurs.[23] As an analogy consider the way in which the visual system somehow constructs or creates the experience of three dimensional objects from more primitive kinds of visual information and various

[23] Similarly for cases in which causal judgment is influenced by covariational information. We can recognize that such influence occurs by manipulating the covariational information and seeing how a subject's judgments vary in response, and we can then try to construct a computational theory connecting the two. This does not require that we must suppose that subjects are consciously carrying out this computation or that they consciously represent the contingency information in a completely 'pre-causal' way.

kinds of complex processing. Most or all of this occurs outside of consciousness. Plainly it would be misguided to criticize accounts of visual object processing on the grounds that we have no phenomenological awareness of this processing or of the more visually primitive elements over which it operates. Similarly for causal learning and inference.

References

Beck, S.R., Riggs, K., & Burns, P. (this volume) 'Multiple Developments in Counterfactual Thinking'.

—— Robinson, E., Carroll, D., & Apperly, I. (2006) 'Children's Thinking About Counterfactuals and Future Hypotheticals as Possibilities', *Child Development* 77: 413–26.

—— Carroll, D.J., Brunsdon, V.A.E., & Gryg, C.K. (2011) 'Supporting Children's Counterfactual Thinking with Alternative Modes of Responding', *Journal of Experimental Child Psychology* 108: 190–202.

Bonawitz, L., Ferranti, D., Saxe R., Gopnik, G., Meltzoff, A., Schulz, L., & Woodward, J. (2009) 'Just do it? Investigating the Gap between Prediction and Action in Toddlers' Causal Inferences', *Cognition* 115: 104–17.

Cheng, P.W. (1997) 'From Covariation to Causation: A Causal Power Theory', *Psychological Review* 104: 367–405.

Collins, J. Hall, N., & Paul, L. (2004) 'Counterfactuals and Causation: History, Problems, and Prospects'. In: *Causation and Counterfactuals*, Collins, J., Hall, N., and Paul, L. (eds), pp. 1–58. Cambridge, MA: MIT Press.

Edgington, D. (this volume) 'Causation First: Why Causation is Prior to Counterfactuals'.

German, T.P. & Nichols, S. (2003) 'Children's Counterfactual Inferences about Long and Short Causal Chains', *Developmental Science* 6: 514–23.

Gopnik, A. (2009) *The Philosophical Baby*. New York: Farrar, Straus, and Giroux.

—— & Schulz, L. (2007) (eds) *Causal Learning: Psychology, Philosophy, and Computation*. New York: Oxford University Press.

Harris, P.L., German, T., & Mills, P. (1996) 'Children's Use of Counterfactual Thinking in Causal Reasoning', *Cognition* 61: 233–59.

Hitchcock, C. (2007) 'Prevention, Preemption, and the Principle of Sufficient Reason', *Philosophical Review* 116: 495–532.

Lewis, D. (1973) 'Causation', *Journal of Philosophy* 70: 556–67.

List, C. & Menzies, P. (2009) 'Non-Reductive Physicalism and the Limits of the Exclusion Principle', *Journal of Philosophy* 106: 475–502.

McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009) 'Cue Competition Effects and Children's Causal and Counterfactual Judgments', *Developmental Psychology* 45: 1563–75.

Mackie, J. (1974) *The Cement of the Universe*. Oxford: Oxford University Press.

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.

—— & Lehman, D.R. (1996) 'Counterfactual Thinking and Ascriptions of Cause and Preventability', *Journal of Personality and Social Psychology* 71: 450–63.

Meltzoff, A. (1988) 'Infant Imitation after a 1-week Delay : Long -term Memory for Novel Acts and Multiple Stimuli', *Developmental Psychology* 24: 470-76.

—— (2007) 'Infant's Causal Learning: Intervention, Observation, Imitation'. In A. Gopnik & L. Schulz (2007), pp. 37–47.

Mitchell, C., Lovibond, P., & Condoleon, M. (2005) 'Evidence for Deductive Reasoning in Blocking of Causal Judgments', *Learning and Motivation* 36: 77–87.

Perner, J. & Rafetseder, E. (this volume) 'Counterfactual and Other Forms of Conditional Reasoning: Children Lost in the Nearest Possible World'.

—— Sprung, M., & Steinkogler, B. (2004) 'Counterfactual Conditionals and False Belief: A Developmental Disassociation', *Cognitive Development* 19: 179–201.

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010) 'Counterfactual Reasoning: Developing a Sense of "Nearest Possible World"', *Child Development* 81: 376–89.

Riggs, K.J., Peterson, D.M., Robinson, E.J., & Mitchell, P. (1998) 'Are Errors in False Belief Tasks Symptomatic of a Broader Difficulty with Counterfactuality?', *Cognitive Development* 13: 73–90.

Robinson, E.J. & Beck, S. (2000) 'What is Difficult about Counterfactual Reasoning?'. In: *Children's Reasoning and the Mind*, P. Mitchell & K.J. Riggs (eds), pp. 101–19. Hove, East Sussex: Psychology Press.

Roessler, J. (this volume) 'Perceptual Causality, Counterfactuals, and Special Causal Concepts'.

Rubin, D. (1974) 'Estimating the Causal Effects of Treatment in Randomized and Non-Randomized Studies', *Journal of Educational Psychology* 66: 688–701.

Saffran, J., Johnson, E., Aslan, R., & Newport, E. (1996) 'Statistical Learning by 8-month Old Infants', *Science* 274: 1926–28.

Schulz, L., Gopnik, A., & Glymour, C. (2007) 'Preschool Children Learn about Causal Structure from Conditional Interventions', *Developmental Science* 10: 322–32.

Shanks, D. (2004) 'Judging Covariation and Causation'. In: *Handbook of Judgment and Decsion-Making*, D. Koehler & N. Harvey (eds). pp. 220–39. Oxford: Blackwell Publishing.

Sobel, D. (this volume) 'Domain-Specific Causal Knowledge and Children's Reasoning about Possibility'.

Ursu, S. & Carter, C. (2005) 'Outcome Representations, Counterfactual Comparisons and the Human Orbitofrontal Cortex: Implications for Neuroimaging Studies of Decision-Making', *Cognitive Brain Research* 23: 51–60.

Woodward, J. (1990) 'Supervenience and Singular Causal Claims'. In: *Explanation and Its Limits* (Royal Institute of Philosophy Conference), Dudley Knowles (ed.), pp. 211–46. Cambridge: Cambridge University Press.

—— (2003) *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

—— (2006) 'Sensitive and Insensitive Causation', *The Philosophical Review* 115: 1–50.

—— (2007) 'Interventionist Theories of Causation in Psychological Perspective'. In A. Gopnik & L. Schulz (2007), pp. 19–36.

—— (2011) 'Causal Perception and Causal Understanding'. In: *Perception, Causation, and Objectivity*, J. Roessler, H. Lerman, & N. Eilan (eds.), pp. 229–63. Oxford: Oxford University Press.

2

The Relationship between Children's Causal and Counterfactual Judgements

Teresa McCormack, Caren Frosch, and Patrick Burns

As should be clear from the contributions to this volume by philosophers, there is a long tradition of philosophical accounts of causation that link causal and counterfactual claims (Collins, Hall, & Paul, 2004; Hart & Honoré, 1985; Lewis, 1973). Put simply, the basic idea is that what it means to say that C causes E is not just to say that if C occurs, E will occur, but also to say that if C were not to occur, E would not occur. Various accounts finesse this general claim in considerable detail (see Collins et al. 2004); the most recent influential related theory is that of Woodward (2003, 2007, this volume). As Woodward points out, most philosophical accounts are 'intended as accounts about the world, rather than accounts of anyone's psychology' (2007: 25). Nevertheless, it may be a worthwhile psychological project to explore the extent to which such theories can inform descriptive accounts of how causal learning and cognition actually function.

From a psychological perspective, we can make a simple distinction between two possible ways in which counterfactual and causal reasoning might be related. The first, which we will call the counterfactual process view, is that counterfactual thought is typically a necessary part of the processing involved in making causal judgements. The alternative possibility is what we will call the psychological relatedness view: that counterfactual judgements are closely related psychologically to counterfactual judgements in so far as (1) the ability to make a causal judgement is usually accompanied by the ability to make the appropriate counterfactual judgement and vice versa, and (2) at least under some circumstances counterfactual and causal judgements are consistent with each other. On the psychological relatedness view, however, counterfactual cognition may not necessarily play a role in the process of reaching causal conclusions. The aim of this chapter is to consider how these viewpoints have influenced research by developmental psychologists, and to discuss whether our own experimental studies provide evidence in support of either of these approaches.

1 Psychological Process View

For developmental psychologists, the most familiar version of a process view comes from Harris, German, and Mills's (1996) influential paper, in which they argued that counterfactual thought may play a central role in children's causal judgements. In this paper, they drew heavily on analyses from the philosophical literature that closely link counterfactual and causal claims, in particular that of Mackie (1974). One of the aims of Harris et al.'s paper was to demonstrate that young pre-schoolers can not only make counterfactual judgements when asked to do so, but will spontaneously invoke counterfactual alternatives to events in situations in which they are asked to explain why certain outcomes had occurred or when asked how such outcomes could have been prevented. Harris et al. concluded that it is likely that children's causal interpretations of events are typically informed by a consideration of counterfactual alternatives, and that this could be the case from very early in development.

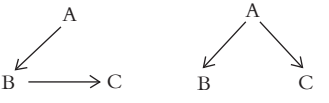
The findings of recent research with adults have brought into question the notion that causal judgements are frequently underpinned by a consideration of counterfactual alternatives (for review, see Mandel, this volume). Furthermore, there is considerable debate over whether Harris et al. (1996) convincingly demonstrated that pre-schoolers are capable of genuine counterfactual thought (see contributions by Beck, Riggs, & Burns, and by Rafetseder & Perner to this volume). It is fair to say that Harris et al.'s suggestion that children's causal judgements are typically arrived at by a process of considering counterfactual alternatives has not been widely adopted by developmentalists. Nevertheless, we can still think about whether it is possible to identify any circumstances in which children's causal judgements might plausibly recruit counterfactual reasoning. In the final section of this chapter, we discuss one possible circumstance in which children may draw conclusions based on counterfactual reasoning, and outline some initial experimental evidence that supports this suggestion. However, we first turn to a consideration of the alternative claim that we have termed a psychological relatedness view: that counterfactual judgements are closely related to causal judgements.

2 Psychological Relatedness Views

One of the most influential recent accounts of causal learning that can be interpreted as a psychological relatedness view has been put forward by the developmental psychologist Alison Gopnik. The causal Bayes net approach espoused by Gopnik and colleagues (e.g. Gopnik et al. 2004; Gopnik, Sobel, Schulz, & Glymour, 2001; Sobel, Tenenbaum, & Gopnik, 2004), has not only generated considerable debate about the nature of causal learning, but has also provided a new context in which the relationship between causal and counterfactual cognition can be considered. This account is a version of the causal models approach to causal learning (Sloman, 2005), since it assumes that both children and adults operate with models of structure of the causal relationships between variables, rather than simply, for example, learning associations

between them. On the causal Bayes net approach, these models are assumed to essentially represent patterns of conditional probabilities between variables, and are constructed and updated in a way that obeys Bayesian principles. For present purposes, it is not necessary to describe the details of such principles (see various contributions to Gopnik & Schulz, 2007, for reviews). Rather, the key relevant assumption of this approach is that such models can be used to predict the effects of hypothetical interventions on (i.e. manipulations to the values of) variables within a given model. In other words, the claim is that such models allow systematic predictions to be made about the likely outcome of altering certain aspects of the world. Because of this property, it has been argued that these inferential structures not only support causal judgements but can also be used to generate counterfactual predictions (Gopnik & Schultz, 2007; see also Hagmayer, Sloman, Lagnado, & Waldmann, 2007; Sloman & Lagnado, 2005). As Hagmayer et al. describe, doing this would require constructing an initial model of the causal relationships between aspects of the world on the basis of observation, and then performing an imaginary selective intervention on the relevant variable within the previously constructed model. Given certain assumptions about the nature of the represented relationships between the variables in the model (in particular, that they are genuinely causal relationships rather than simply probabilistic ones), accurate counterfactual prediction should be possible. Thus, the causal Bayes net framework seems to readily yield an interpretation in terms of what we have called a psychological relatedness approach.

Thus, this general approach makes a clear empirical prediction: if participants have extracted a causal model on the basis of their observations of variables, they should be able to make appropriate counterfactual predictions. To illustrate, consider a situation in which a participant is asked to make counterfactual predictions about the effects of intervening on a variable after extracting either a common cause causal model or a causal chain model of the relationships between three variables (see Figure 2.1). In a common cause model, variable A is represented as the common single cause of B and C, whereas in a causal chain model, A is represented as causing an intervening event B which then causes a final effect C. A crucial difference between these models is what they should predict about the effect of making a hypothetical intervention on variable B. For example, other things held constant, preventing B from occurring should have no effect on whether C occurs following A under a common cause model, whereas under a causal chain model, C should not be able to occur if B is prevented from occurring. Thus, this type of account assumes that participants should vary their



Causal Chain Common Cause

Figure 2.1 Causal chain and common cause structures.

predictions about the effects of preventing the occurrence of B depending on whether they have been given observations that lead them to extract a causal chain versus a common cause model. Below, we discuss some of our experimental studies that have examined whether or not this is the case in children.

We note that Gopnik and colleagues do not in their writings assume that a process account of causal judgement holds (i.e. that causal conclusions are typically reached by a process of considering counterfactual alternatives). Indeed, Schulz, Kushnir, & Gopnik (2007) accept that children can acquire causal knowledge by many routes. However, at least in some writings, it would appear that they intend their account to go further than the claim that causal representations should support counterfactual judgements. Specifically, influenced by the interventionist approach to causation in philosophy (Woodward, 2003, 2007, this volume), Schulz, Kushnir, & Gopnik (2007) suggest that causal knowledge is in fact essentially knowledge about the effects of observed, hypothetical, or counterfactual interventions. That is, roughly speaking, what it is to represent A as the cause of B just is a matter of being committed to the idea that intervening on or manipulating the value of A will affect B. They contrast such an approach with more widely accepted approaches within developmental psychology that characterize causal knowledge in terms of the knowledge about mechanism (e.g. most notably, that of Shultz, 1982; but see also Schlottmann, 1999, 2000; White, 1995). Although Schulz, Kushnir, & Gopnik acknowledge that children may possess knowledge about the operation of specific mechanisms that they have encountered, such mechanism knowledge is not thought to be basic to representations of causal relationships. Thus, Schulz, Kushnir, & Gopnik argue that 'a causal relation is defined . . . in terms of the real and counterfactual interventions it supports' (2007: 69), and that 'when children infer that a relationship is causal, they commit to the idea that certain patterns of interventions and outcomes will hold' (2007: 70). Our experimental studies examined whether this is indeed the case.

3 The Consistency of Children's Causal and Counterfactual Judgements: Causal Structure Judgements

In a series of studies, we have examined whether Schulz, Kushnir, & Gopnik (2007) are correct in suggesting that children's judgements about the effects of counterfactual interventions should be consistent with their causal judgements. In all of these studies, we have examined scenarios in which children extract the causal structure of a three-variable system as either a causal chain or a common cause, and are then asked to make explicit judgements about the effects of intervening on one or more of the variables. In one study (Frosch, McCormack, Lagnado, & Burns, in press), children were provided with very simple temporal cues about causal structure. Children observed a box with three differently coloured components on top of it that moved circularly on the

horizontal plane, which for present purposes we can label A, B, and C. The temporal structure of the events involving A, B, and C varied between trials, and children saw a number of demonstrations of the event series in each trial. The components were spatially separated with no obvious mechanisms connecting them; in fact, unbeknownst to participants, they were controlled by a computer hidden inside the apparatus. In *sequential* trials, A moved, followed by a short delay, then B moved, followed by an additional short delay, and then C moved. In this condition, children as young as 5 years judged that A, B, and C formed a causal chain, by selecting an appropriate diagram to match their model choice. Following Schulz, Gopnik, & Glymour (2007), these diagrams were in the form of anthropomorphized pictures in which hands reaching from one component to another signified a causal link (see Figure 2.2a for an example; the particular objects used as A, B, and C varied between trials). In *synchronous* trials, A moved, and then after a short delay B and C moved simultaneously. In these trials, 5-year-olds reliably judged that A was a common cause of B and C (see Figure 2.2b for an example). After children had selected their model choices, we placed the diagram of the model in front of them so that they could remember what causal

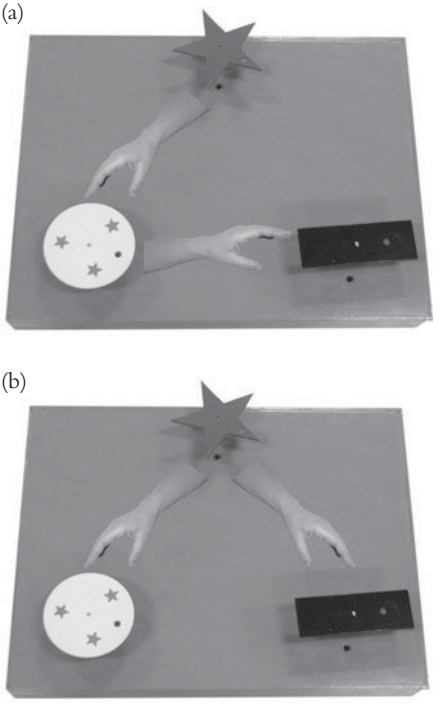


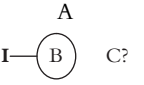
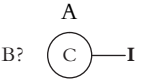
Figure 2.2 (a) Diagram of test apparatus showing a causal chain structure; (b) Diagram of test apparatus showing a common cause structure.

structure they had selected. Given that the only difference between the sequences that children observed across trial types was the temporal arrangement of the events, we can be confident that the cross-trial differences in children's causal structure judgements were a result of these judgements being based on the simple temporal cues provided.

After children had made their causal structure judgements, we then asked them specific counterfactual questions, the answers to which should have differed across the trial types. Children were asked what would have happened during the previous demonstrations of the box's operation if B or C had been unable to operate. More specifically, they were asked whether C would have moved if B had been unable to operate, and whether B would have moved if C had been unable to operate. We showed children how individual components could have been disabled (described to children as 'stopped from working') using a small vertical bar (fashioned as a 'Stop' sign) that prevented a component from moving. If children were correctly using the causal models that they had extracted to make counterfactual judgements, their answers to these questions should have differed appropriately across trial types (see Table 2.1 for correct answers for each model). In particular, when children had judged the structure to be a causal chain structure, they should have predicted that C would not have moved if B had been stopped from working. However, this was not the case. Children's counterfactual judgements had low consistency with their causal model choices, and did not differ significantly across trial types. This was true for older children (6–7-year-olds) as well as 5-year-olds, making it unlikely that children's difficulties stemmed from understanding the counterfactual questions. Thus, the findings of this study do not support the suggestion that children's representations of causal structure support counterfactual judgements about interventions appropriately.

Although we do not believe that children's difficulties stemmed from understanding the counterfactual questions that we posed, evidence from some developmental studies (e.g. Perner, Sprung, & Steinkogler, 2004; Riggs, Peterson, Robinson, & Mitchell, 1998) suggests that young children may find questions posed as future hypotheticals (questions of the form 'what will happen') somewhat easier than counterfactual questions (questions of the form 'what would have happened'). This issue is discussed

Table 2.1 Correct answers to the counterfactual questions about the effect of intervening on either B or C on the operation of the other component

	Question Type: Would C/B have moved?	
		
Common Cause	Yes	Yes
Causal Chain	No	Yes

Note: **I** indicates an intervention on the circled component which disables it.

in Perner and Rafetseder's contribution to this volume. Indeed, Woodward (this volume) argues that an interventionist account of causal cognition need only assume that causal representations support judgements about future hypothetical circumstances, rather than those which psychologists would usually consider to be counterfactual. In another study (Burns & McCormack, 2009) we used a similar procedure, with children extracting causal structure based on temporal cues for three variables, A, B, and C, again as either causal chain or common cause structures. Children were asked sets of causal questions ('which one made B go?') and also asked to choose a diagram that reflected their causal judgements. Following these judgements, children were asked to make judgements about interventions. The experimenter carried out the intervention by physically preventing either component B or C from operating, and then asked children a future hypothetical question about whether the other component (B or C) would operate when the event sequence was initiated. In this study, children were shown only a single trial type to ensure that their answers were not affected by observing other causal structures. Even under these circumstances, 6–7-year-olds' judgements about interventions were not consistent with their causal model judgements. Adults were tested as well, and we found that their intervention judgements were consistent with their causal model judgements, and varied appropriate across trial types.

One possible explanation of these findings is that the judgements children have to make have two components, in so far as they involve inferring the consequences of (i) preventing one event (B or C) from occurring and then (ii) the subsequent generation of the overall event sequence by initiating event A. We label such interventions as prevent-then-generate interventions, because of the requirement to infer the consequences of both of these components. Arguably, such interventions are potentially more complex to reason about than interventions that simply involve inferring the consequences of manipulating a single component (e.g. what would happen if B or C themselves were manipulated?), which we refer to as simple generative intervention judgements. In a further study, children were asked to make a more simple intervention judgement. After they had selected their causal model they were asked a future hypothetical question about what would happen to either B or C if the other component were made to move (e.g. 'if I move B like this, will C go?'). Again, 6–7-year-olds' answers to these more simple generative intervention questions were not consistent with their causal structure judgements and did not differ appropriately across trial types.

Thus, across a series of studies we did not find support for Schulz, Kushnir, & Gopnik (2007) claim that 'when children infer that a relationship is causal, they commit to the idea that certain patterns of interventions and outcomes will hold' (2007: 70). Although children inferred that they were observing different causal structures across the two trial types, their counterfactual and future hypothetical judgements about the effects of interventions were not consistent with these inferences. The causal structures that we used were very simple, and indeed they were chosen because they are generally

considered to be the building blocks of more complex causal structures (along with a three-variable common effect model; Hagmayer et al. 2007). However, we note that our findings are potentially inconsistent with those of Schulz, Gopnik, & Glymour (2007, Exp. 2). In their study, children were told that a system of two gears had a particular causal structure, illustrated using diagrams similar to those that we used. They then asked children whether each one of the gears would move when the other one was removed. They found that, across four different causal structures, children answered these questions correctly more frequently than would have been expected by chance. We note, though, that although performance was above chance, in no instance did the majority of children give the correct answers to these hypothetical questions (i.e. no more than eight out of their sample of sixteen children answered the questions correctly for any given causal structure). Nevertheless, the authors argue that the findings suggest that children can use their causal structure knowledge to make predictions about the effects of interventions. There are a large number of differences between our methodology and that of Schulz, Gopnik, & Glymour (2007), including the physics of the apparatus and the number of moving components (e.g. they used two interlocking gears whereas we used three spatially distinct components). Here we focus on just one difference that may potentially be important. Whereas in our studies children were given temporal cues to derive causal structure, Schulz Gopnik, & Glymour (2007, Exp. 2) simply told children what the causal structure was.

The fact that children had to initially infer a causal structure based on temporal cues raises two possible interpretations of our results. On the first interpretation, it could be argued that although children appeared to infer causal structure appropriately based on temporal cues, it is possible that their causal model choices simply directly reflected their perceptions of the temporal relationships between the events, rather than function as true reflections of how they represented the causal relationships between the events. In other words, children chose a diagram that matched the temporal pattern of events and did not interpret this pattern in causal terms. If this first interpretation of our findings is correct, then it is not surprising that we found that children had difficulty answering intervention questions. The obvious way to address this possible is to follow Schulz, Gopnik, & Glymour (2007) in simply telling children what the causal structure is, rather than children having to infer it, before asking intervention questions.

A second possible interpretation of the findings is that children did indeed use the temporal information to infer a truly causal structure, but that some kinds of inferences, including those based on temporal information as in our study, do not necessarily support intervention judgements. In other words, whether or not causal structure representations reliably support intervention judgements maybe highly dependent on the basis on which the causal representations have been made. We know of no evidence to directly support this suggestion. One potentially related finding comes from Lagnado and Sloman's (2004, Exp. 4) study, in which they examined the role of temporal versus covariation cues in causal structure judgements. They found that if participants selected a causal model based on temporal cues, the judgements they made

about the probabilistic dependencies that held between events were not consistent with their causal model choices. Although these judgements were not intervention judgements per se, this finding supports the suggestion that judgements based on temporal information may support some types of causal judgements but not others.

In this context, it is worth considering whether the causal judgements in our task may be more appropriately seen as akin to the sort of judgements made in perceptual causation tasks rather than causal reasoning tasks (the latter typically involve participants making an inference based on patterns of covariation information; e.g. Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). The temporal delays between the events were brief, and indeed it may be more correct to describe participants as perceiving that (e.g.) A caused B which then caused C rather than inferring such a causal structure. Perhaps if causal structure judgements are less cognitively mediated they are less likely to support intervention judgements, and this may be particularly the case in children. Although adults made intervention judgements that were consistent with their causal structure judgements, this may have been because they recognized the need for such consistency when provided with a reminder of their structure judgement (the diagram they had chosen), whereas children may not have recognized the need for such consistency. The suggestion that we should distinguish, broadly speaking, between perceptual versus causal judgements based on inference is a long-standing one within the psychological literature (e.g. Roser, Fugelsang, Dunbar, Corballis, & Gazzaniga, 2005; Schlottmann & Shanks, 1992) and it would be interesting to explore the implications of such a distinction for intervention judgements.

4 The Consistency of Children's Causal and Counterfactual Judgements: Causal Power Judgements

In our studies reviewed here, we invariably found that children's causal structure judgements were not reliably consistent with their hypothetical or counterfactual intervention judgements. However, we have also examined the relationship between children's causal and counterfactual judgements in a circumstance in which the children's task was not to extract the causal structure of the relationships between variables, but to infer whether or not a specific object possessed a causal power. In this context, we have found more evidence to suggest a relationship between causal and counterfactual judgements.

Historically, much research on the development of causal cognition has focused on children's ability to detect and make use of information about the causal powers of novel objects, rather than to make causal structure judgements.[1] It has been well established that children can learn about the causal powers of novel objects quickly

[1] The term 'causal power' is used loosely here without implying any commitment to any particular theory of causal learning or knowledge, or indeed the metaphysics of causation.

and in apparently principled ways (Shultz, 1982; Sedlak & Kurtz, 1981). Indeed, much of the debate over the nature and development of causal reasoning concerns the type of principles underpinning such learning (White, 1995). The most recent series of studies examining children's ability to learn about causal powers have typically used the 'blicket detector' task, in which children observe objects being placed on a novel machine either individually or in pairs (e.g. Gopnik et al. 2001; Kushnir & Gopnik, 2007; Sobel, Tenenbaum, & Gopnik, 2004; Sobel & Kirkham, 2006). Objects that are causally efficacious and make the machine operate (light up and play a tune) are labelled 'blickets', and children's task is to judge whether particular objects are or are not blickets.

In a series of studies using the blicket detector task, we have examined children's causal and counterfactual judgements, with the aim of exploring whether these types of judgements show similar developmental patterns (McCormack, Butterfill, Hoerl, & Burns, 2009). The nature of the task enabled us to ask either causal or counterfactual questions to tap causal knowledge. Because objects can be placed on the detector not just individually but in pairs, it is possible to ask children not only if a particular object is a blicket (a causal question), but also whether the detector would have gone off in the absence of one of the objects (a counterfactual question). The fact that objects can be placed on the detector either individually or along with other objects makes it possible to use this paradigm to examine *cue competition effects*. These are effects that demonstrate that whether or not a cue A is judged to be causally efficacious in bringing about an outcome is not just a matter of the extent to which the outcome co-occurs with A but the extent to which other competing cues are also related to the outcome. Such effects have been the touchstone for the development of theories of causal learning (e.g. Aitken & Dickinson, 2005; Shanks, 2007). The most well-known effect is that of blocking. In blocking tasks, participants (animal or human) are shown that one stimulus is predictive of an outcome and then shown that when this stimulus is presented simultaneously along with another new stimulus, the outcome also occurs. For example, a frequently used task with adults is one in which participants have to find out which foods cause an allergic reaction in a fictional patient. Participants are initially shown that the reaction occurs when one food, Food A (e.g. nuts), is eaten on its own. They are then shown that the reaction also occurs when the patient eats an additional food, Food B (e.g. cheese), simultaneously with Food A, which they had seen previously paired with the allergic reaction. The standard finding is that having been shown that the reaction occurs when (e.g.) nuts are eaten on their own blocks any new learning that (e.g.) cheese also causes the allergic reaction.

Thus, in blocking procedures, participants initially learn about a cue A that is paired with an outcome (i.e. they see A+ trials; element training). In the case of the blicket detector task, this cue is a specific object that sets off the blicket detector. They are then shown a compound cue AB, also paired with the outcome (i.e. AB+; compound training). In the blicket detector task, the compound cue is a pair of objects placed simultaneously on the detector, and the detector goes off in the presence of this pair of

objects. As we have said, exposure to A+ trials typically *blocks* new learning of the pairing of B with the outcome, with ratings of B's causal efficacy lower than if A+ trials had not been demonstrated independently (e.g. Chapman & Robbins, 1990; Le Pelley, Oakeshott, & McLaren, 2005; Shanks, 1985). Thus, in our task, participants should be less likely to judge that B is a blicket if they have only seen it placed on the detector as part of a pair, along with A that has already been shown to set off the detector by itself. A related effect is that of *prevention from overshadowing*, in which A is independently demonstrated not to be paired with the outcome (i.e. A- trials). Under such circumstances, showing AB+ trials leads to particularly strong ratings of B's causal efficacy. Thus, if participants see two objects A and B placed on the detector together, and the detector goes off, they should be particularly likely to judge that B is a blicket if they have already been shown separately that A does not set the detector off. In all of these procedures, causal ratings to B should be compared to ratings for a cue from a control compound in which neither cue is demonstrated individually (De Houwer, Beckers, & Vandorpe, 2005). Indeed we used such control trials in our studies. Participants were shown two novel objects placed on the detector as a pair, with neither object demonstrated individually, and asked to judge whether one item from this control pair was a blicket. Table 2.2 shows the typical presentations given in each different type of trial in our studies.

In the original versions of these procedures, element trials, i.e. the trials involving A+/- presentations, were given before the compound AB+ trials. However, it has subsequently been found that if A+/- trials are given after AB+ trials, under at least some circumstances B's status is retrospectively revaluated, giving the parallel effects of *backward blocking* and *release from overshadowing* (e.g. Chapman, 1991; Kruschke & Blair, 2000; Wasserman & Berglan, 1998). The theoretical significance for theories of causal learning of such retrospective learning has been much discussed (e.g. Denniston, Savastano, & Miller, 2001; Dickinson & Burke, 1996; Kruschke, 2006). In our studies, we used both backward and forward versions of these procedures, but as we did

Table 2.2 Presentations of element and compound stimuli used to demonstrate various cue competition effects

	Phase 1 (elements)	Phase 2 (compound)
Forward blocking	A+	AB+
Prevention of overshadowing	A-	AB+
Backward blocking	AB+	A+
Release from overshadowing	AB+	A-
Control pair	CD+ (backward designs)	CD+ (forward designs)

Note: + indicates that the detector went off when the object was placed on it, - indicates that the detector did not go off. The compound stimuli involve placing both objects on the detector together.

Table 2.3 Sample parallel counterfactual questions to the causal questions

	Phase 1	Phase 2	Causal Question: Is B a blicket?	Counterfactual Question: Would the machine have gone off without A?
Backward blocking	AB+	A+	If blocking, 'No'	If blocking, 'No'
Release from overshadowing	AB+	A-	Yes	Yes

not find any difference between these conditions, we will not distinguish between them in this discussion of our findings.

In our initial experiment, we asked children simply to make causal judgements regarding B's status (i.e. is B a blicket?). We found a clear developmental pattern with regard to the set of cue competition effects that we examined when children were asked to make these causal judgements. There was no evidence of either blocking or prevention/release from overshadowing in our 4-year-old age group, but our 5–6-year-old group showed significant levels of all of these effects.[2] Thus, we were confident that by the time children were 5–6 years old, their causal judgements reflected the patterns of presentations of cues that they had seen. Their judgements about the causal status of B took into account not just how often the detector went off when B was placed on it, but the status of the A objects that always accompanied the B objects. This task then provided us with an interesting context in which to examine children's counterfactual judgements. The structure of the trials is such that, although they have to make a causal judgement regarding B, participants never see B presented on its own. Essentially, in test trials they are only provided with indirect evidence about B's status in so far as they must take into account their observations of whether or not the detector went off when A was presented on its own. The implication of this is that there are counterfactual questions that have natural and obvious parallels to the causal questions, namely whether the detector would have gone off if only B had been present. Thus, if participants have judged that B is *not* a blicket, we would expect them to also judge that the detector would not have gone off if B had been placed on it without A. Similarly, if they have judged that B *is* a blicket, they should also judge that the detector would have gone off if B had been placed on it without A. Table 2.3 illustrates these sample questions and predicted answers.

In a follow-up series of studies, we used a similar procedure to our initial study in terms of trial structure, but rather than asking children whether or not B was a blicket, we asked them the parallel counterfactual questions: whether the detector would have gone if A had not been placed on it along with B. For example, in one study, following

[2] The age differences that we found do not match those of a previous related study, that of Sobel et al. (2004). However, the Sobel et al. study did not use the same control trials as our study; see McCormack et al. (2009) for detailed discussion.

the last presentation, the experimenter said to the children 'A moment ago I put both of these blocks on the detector together and it went off. Do you think it would have gone off if I hadn't put this one (gave colour of A) on?' (see McCormack et al. 2009 for further details). Strikingly, we found an identical pattern of age effects when children were asked these parallel counterfactual questions as when children were asked the causal questions. Again, 4-year-olds showed no evidence of cue competition effects, but these were in evidence for 5–6-year-olds, with the older children showing the appropriate differences between experimental and control questions in their answers to counterfactual questions. Thus, older children were capable of making counterfactual judgements, and age differences were similar to those observed when the parallel causal questions had been asked.

These counterfactual questions are interesting, because they involve children considering a counterfactual scenario that they have never observed, in the context of a task involving a novel causal property and a set of unfamiliar objects. Thus, it would not be possible for children to answer the counterfactual questions correctly on the basis of some type of general knowledge: answering correctly depended on considering what had actually happened in the previous demonstrations. We would argue that our findings provide convincing evidence that by at least this age children are capable of genuinely counterfactual judgements, because children could not have answered the questions without (i) considering what had actually happened and (ii) considering a never-observed scenario (see Perner & Rafetseder, this volume, for discussion of these issues).

Note that we would not want to conclude from these findings that our 4-year-olds were not capable of counterfactual judgements. Rather, we believe that it is more likely that they understood the counterfactual questions, but failed to use the patterns of observations they were provided with to draw the appropriate conclusions, just as they failed to do so when asked causal questions (whether or not B was a blicket). Some unpublished data support this suggestion. In a follow-up experiment, children were again asked counterfactual questions, but this time they were asked about what would have happened if B had not been placed on the detector along with A (i.e. if A had been placed on the detector on its own). In other words, they were asked the counterfactual parallel to the causal question, 'Is A a blicket?'. In our initial study with causal questions, even 4-year-olds were virtually at ceiling in judging whether or not A was a blicket, with no age differences in causal judgements, since children had actually seen what happened when A was placed on the detector on its own (they had direct evidence about A's causal status). That is, answering the counterfactual question in this instance does not require children to infer across patterns of evidence, but simply to consider the observations they have seen of A on the detector on its own. We found that 4-year-olds, while not at ceiling in answering the counterfactual questions, showed appropriate patterns of responses across the different trial types, and were in fact as accurate as 5–6-year-olds in this task. Thus, it seems to us to be likely that 4-year-olds can at least make sense of counterfactual questions in this type of task.

When they have been provided with direct evidence about a cue's causal status, they are as good as older children in answering a counterfactual question that involves sensitivity to the status of that cue. What may differ with age is the ability to make counterfactual judgements that take into account the indirect evidence children have been presented with about a cue's status.

What are the implications of our findings for what we have described as a psychological relatedness view? It is clear from our findings that the learning mechanisms and representations that children use to support their causal judgements in this task also support counterfactual judgements. This is, at a minimum, the type of theoretical commitment that a psychological relatedness view would hold. However, it is also clear that the results described in our previous section on children's causal structure judgements are not supportive of this claim. Thus, we need to consider why there might be a difference between the two sets of findings. The most salient difference, of course, lies in the nature of the causal knowledge that the two tasks assess, with one task requiring children to consider the impact of a counterfactual or hypothetical intervention on a three-variable system, which involves considering all the causal dependencies of variables, and the other requiring children only to consider in binary terms whether or not an object possess a causal power. Arguably, the first sort of reasoning is more complex than the second.

Another possibility, mentioned above, is that the consistency of children's causal and counterfactual judgements may depend on the type of information used to extract causal knowledge. With regard to our causal structure studies, we suggested that the fact that children's judgements were made on the basis of simple temporal cues may have meant that these judgements were somehow less cognitively mediated than judgements that might have been made on the basis of some other type of information (e.g. information about the effects of actual interventions, or covariation information). We acknowledged that this is speculative, but nevertheless one version of this speculative proposal may explain the difference between our findings across these two types of paradigms in an interesting way. The suggestion we wish to consider is that counterfactual thought *may already be involved* in the process of reaching appropriate causal conclusions in our blicket detector task, whereas this is not the case in our causal structure task, in which causal judgements reflected the use of temporal cues. If making causal judgements already involves thinking counterfactually in our version of the blicket detector task, it would not be surprising that both causal and counterfactual questions show similar developmental patterns.

5 Counterfactual Reasoning and Blocking

In this section, we consider the possibility that counterfactual reasoning may be part of the process of making causal judgements in our blocking task. Blocking was first demonstrated in animals, and indeed some highly influential theories of causal learning have their roots in accounts of animal learning, most notably accounts that describe

such learning in terms of the formation of associations between stimuli and outcomes (for review, see De Houwer & Beckers, 2002; Dickinson, 2001; Shanks, 2007). By contrast, some recent approaches to human causal learning emphasize quite different types of processes, in particular the role of explicit reasoning processes in making causal judgements (e.g. De Houwer & Beckers, 2002; De Houwer, Beckers, Vandorpe, 2005; Lovibond, Been, Mitchell, Bouton, & Frohardt, 2003). Although it may be usual to accord such processes a role in the types of tasks frequently studied within the realm of social psychology (e.g. tasks in which participants have to judge which of several characters should be held responsible for an unpleasant event, such as those discussed by Mandel, this volume, and Hitchcock, this volume), it is only in recent years that the role of such processes has been seriously considered within the context of more traditional causal learning tasks (see Pineño & Miller, 2007). The claim that we wish to consider is that reasoning with a counterfactual component may play a role in explaining blocking phenomena.

In some blocking tasks, the effect is *additive*. If two or more causally efficacious stimuli are present, then the effect is greater than if one only was present (e.g. in a food allergy task, the magnitude of an allergic reaction is greater if two allergy-causing foods are eaten than if just one is consumed). Mitchell, Lovibond, & Condoleon (2005) have pointed out that under such circumstances a chain of reasoning that has a counterfactual component might underpin blocking effects (counterfactual components are given in italics below). So, for example, a participant in a food allergy task might reason 'Food A causes the allergic reaction, and when Food A and Food B were eaten together the allergic reaction was of the same magnitude. *If both Food A and Food B had been causal, the allergic reaction would have been greater than it was*. Therefore, Food B does not cause the reaction.' Note that this specific chain of reasoning only holds when effects are additive. When they are not, the parallel chain of reasoning is: 'Food A causes the allergic reaction. There was also an allergic reaction when Food A and Food B were eaten together. *However, if Food B had not have been eaten along with Food A, there would still have been an allergic reaction*.' This chain of reasoning only licenses the weaker conclusion that one has been provided with no useful information about the causal status of Food B, whereas in the additive case the chain of reasoning licences the much stronger conclusion that Food B cannot be causal. Indeed, Lovibond and colleagues have demonstrated that blocking effects are considerably stronger when effects are additive (Lovibond et al. 2003; see also Beckers, De Houwer, Pineño, & Miller, 2005). Nevertheless, even in non-additive circumstances, if participants use the chain of reasoning described above effectively, we should predict that their judgements about the causal status of B should differ from their judgements regarding the status of stimuli presented in causally informative demonstrations. Thus, even in the non-additive case, a chain of reasoning with a counterfactual component may play a role in causal judgements, since counterfactual reasoning licences the conclusion that essentially no information has been provided about B's causal status. Under such circumstances,

participants should assume that B is only as likely to be a blicket as any object for which they have received no information.[3]

A number of findings in the literature suggest that participants may indeed engage in explicit reasoning in causal learning tasks (e.g. De Houwer & Beckers, 2003; Vandorpe, De Houwer, & Beckers, 2005; see De Houwer et al. 2005, for review). However, as things stand, there is no direct evidence that participants recruit a chain of reasoning with a specifically counterfactual component when they are completing blocking tasks. As we have already described, in our original study we found evidence for blocking in 5–6-year-olds, but no evidence of blocking in a 4-year-old sample (but see Beckers, Van den Broeck, Renne, Vandorpe, De Houwer, & Eelen 2005). The activation of the detector is identical if one or two blickets are placed on it, so this task is an example of a task in which the effect is non-additive. If counterfactual reasoning plays a role in blocking in this task, then the necessary chain of reasoning in this task would be along the following lines: 'A set the detector off on its own. When A and B were placed on the detector together it went off. *But it would have gone off anyway if B had not been there.* Therefore, no information about B's causal status has been provided.' One possible explanation of the age difference that we found was that the younger children did not spontaneously exploit this chain of reasoning that would have led to blocking. Our suggestion is that perhaps younger children are less likely to spontaneously consider a relevant counterfactual alternative when asked to make a causal judgement, and that failure to do so may mean that they are less likely to show blocking on a task in which such a consideration underpins the blocking effect.

To examine this issue, in a follow-up study to that of McCormack et al. (2009), we used the same paradigm but actually provided children with the necessary pieces of counterfactual information that might assist them in making causal judgements. We reasoned that if age differences are reduced or abolished under such circumstances, this might suggest that the developmental patterns that we had found previously might be a result of the differential tendency of the age groups to spontaneously recruit counterfactual reasoning in their causal judgements. Thus, the experimenter provided children with a counterfactual cue before asking the usual causal question (i.e. before asking

[3] Recall that the measure of blocking is given by the difference between causal judgements given to B after both A+ and AB+ demonstrations in comparison to judgements given to C after just CD+ demonstrations (see Table 2.2). Normatively, the likelihood that B is causal under such circumstances is identical to the baseline probability that any given cue is causal (see also Sobel et al. 2004; Sobel & Kirkham, 2006), whereas the likelihood that C is causal will be higher than the baseline unless the baseline probability is 1. Of course, the probability that C is causal is also related to the baseline probability, and can be derived from Bayes law to be 1/(2 – baseline). Thus, for example, if the baseline probability is .50, then the probability that C is causal is .66. If participants are sensitive to these differing probabilities, then we would still expect to see a difference in judgements to B and C even in non-additive conditions. The occurrence of blocking in non-additive conditions is not incompatible with Mitchell et al.'s (2005) account, although it requires an extension of it. In additive conditions, blocking could occur purely on the basis of a logical deduction, as Mitchell et al. (2005) have described. If it is observed in non-additive cases, then it might suggest that participants are, in addition to recruiting an inferential reasoning with a counterfactual component, in some way sensitive to probabilistic information which leads them to give judgements to B that approximate to the baseline.

whether or not B is a blicket). In blocking trials, after all the demonstrations involving the objects had been shown, the experimenter lifted both blocks A and B and said, 'A moment ago I put these two blocks on the detector together and it went off. Actually, it would have gone off anyway if I had only put this (gave colour of A) one on.' Following this, the experimenter then asked about the causal status of B: 'Do you think that this (gave colour of B) one is a blicket?'

Both the 4-year-old group and the older children showed evidence of blocking in this version of the task, and no age differences were found in level of the demonstrated blocking. The pattern of results found in the older age group was very similar to that found in our previous experiments (McCormack et al. 2009). However, the results from the 4-year-olds contrast strikingly with those from our previous study, suggesting that providing the counterfactual cue assisted these younger children in making their causal judgements. The results suggest that children of this age group can understand counterfactual statements and draw upon them in making causal inferences.

Our motivation for conducting this study was to investigate further the possibility that counterfactual reasoning may play a role in explaining blocking phenomena. The findings reported here can be interpreted as very preliminary evidence of such a role, since they suggest that the absence of blocking that we have previously observed in a 4-year-old group may have been due to a failure of children of this age to spontaneously consider counterfactual alternatives. When the requirement to generate such a counterfactual alternative is removed, blocking is observed in this group. We are aware that a role for counterfactual reasoning needs further investigation in both child and adult samples. It remains possible that counterfactual reasoning does not normally underpin blocking in either children or adults, but that when provided with relevant counterfactual information children can nevertheless exploit it. For example, it may be that counterfactual processing is normally only implicit in this task even in older children (see Woodward, this volume), but when young children are made aware of the necessary pieces of counterfactual information by the experimenter they can use them to inform their causal judgements.

It is important to emphasize also that we are not arguing that counterfactual reasoning is likely to underpin blocking phenomena in the various contexts in which they have been studied. Widespread observations of blocking in the animal kingdom indicate that this is highly unlikely. Indeed other studies using quite different procedures have found evidence of blocking in children as young as or younger than the 4-year-old group in our study (Beckers et al. 2005; Sobel & Kirkham, 2006). Rather, the suggestion is that counterfactual reasoning may play a role in some types of tasks that have been used to study human causal learning. In particular, some type of explicit reasoning process that may have a counterfactual component may be more likely to be employed in tasks, such as the present one, which are distinctive in so far as they involve a small number of stimuli/demonstrations. Under these circumstances, it might be particularly likely that explicit reasoning processes play a role because it is relatively easy to recall what happened on each demonstration involving a given object.

As Vandorpe, De Houwer, & Beckers, 2007 point out, in these conditions 'participants can easily keep track of the different events and have ample opportunity to engage in controlled reasoning' (2007: 358). In scenarios in which participants keep track of many stimuli and view dozens of presentations, quite different processes are likely to be important (though see Vandorpe et al. (2007) for discussion). As many theorists have suggested, the specific processes underpinning causal learning are likely to vary depending on the task (e.g. see Lovibond et al. 2003; Pineño & Miller, 2007; Shanks, 2007). However, given the long-standing interest in the relationship between causal and counterfactual judgements, it may be useful to attempt to identify those contexts in which counterfactual reasoning may play a role.

More generally, developmental research may be a particularly fruitful context in which to study the role of explicit inferential reasoning processes, such as counterfactual reasoning, in causal learning. If such processes play a role in explaining certain phenomena in human causal cognition, as has been argued elsewhere (see De Houwer et al. 2005), then we might expect to see developmental changes in the occurrences of such phenomena if the reasoning skills in question are ones which are likely to change developmentally. Furthermore, such accounts might provide a way of explaining developmental changes in causal learning that are not so readily explicable by accounts that focus on lower-level processes.

6 Concluding Thoughts

There is a long-standing research tradition examining children's causal cognition, and a surge of interest over the last decade in examining the development of counterfactual reasoning. However, despite Harris et al.'s (1996) seminal paper, there has been relatively little research that has examined the relationship between these two types of abilities in young children. Developmental research is likely to be particularly important in clarifying the relationship between counterfactual and causal reasoning, not least because intuitively it seems likely that there is an age before which children are not capable of thinking counterfactually. Our own research has only begun to address issues about the relationship between these two types of abilities. As they stand, the findings suggest that much may depend on the nature of the tasks that are used, and the type of causal knowledge involved. We also have some preliminary evidence that suggests that counterfactual reasoning may actually play a role in the process of making some types of causal judgements (in our blocking task). Clearly, there is much scope for additional research to clarify these issues.

We have characterized the minimal theoretical commitment of a psychological relatedness view that the sort of representations that support causal judgements should also support counterfactual ones. A stronger version of this claim is that in essence causal knowledge just is knowledge about the effects of actual, counterfactual, or hypothetical interventions (Schulz, Gopnik, et al. 2007). Arguably, it is this claim about what it is to have causal concepts that provides the clearest theoretical challenge

to developmental psychologists, who historically have tended to conceive of causal concepts in mechanistic terms. It is exactly the sort of claim that developmental research is more likely to be able to address than the sort of research that has been done with adults, because of the possibility of looking at the developmental relationships between the relevant abilities/concepts. As far as we can see, existing research, including our own, has not yet resolved this issue, and we hope that the various contributions to this volume will stimulate the necessary empirical research to address it.[4]

References

Aitken, M.R. & Dickinson, A. (2005) 'Simulations of a modified SOP model applied to retrospective revaluation of human causal learning', *Learning & Behavior* 33: 147–59.

Beckers, T., De Houwer, J., Pineño, O., & Miller, R.R. (2005) 'Outcome additivity and outcome maximality influence cue competition in human causal learning', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31: 238–49.

—— Van den Broeck, U., Renne, M., Vandorpe, S., De Houwer, J., & Eelen, P. (2005) 'Blocking is sensitive to causal structure in 4-year-old and 8-year-old children', *Experimental Psychology* 52: 264–71.

Burns, P. & McCormack, T. (2009) 'Temporal information and children's and adults' causal inferences', *Thinking & Reasoning* 15: 167–96.

Chapman, G.B. (1991) 'Trial order affects cue interaction in contingency judgment', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 17: 837–54.

—— & Robbins, S.J. (1990) 'Cue interaction in human contingency judgment', *Memory & Cognition* 18: 537–45.

Collins, J., Hall, N., & Paul, L.A. (2004) *Causation and counterfactuals*. Cambridge, MA: MIT Press.

De Houwer, J. & Beckers, T. (2002) 'A review of recent developments in research and theories on human contingency learning', *Quarterly Journal of Experimental Psychology* 55B: 289–310.

—— —— (2003) 'Secondary task difficulty modulates forward blocking in human contingency learning', *Quarterly Journal of Experimental Psychology* 56B: 345–57.

—— —— & Vandorpe, S. (2005) 'Evidence for the role of higher-order processes in cue competition and other learning phenomena', *Learning & Behavior* 33: 239–49.

Denniston, J.C., Savastano, H.I., & Miller, R.R. (2001) 'The extended comparator hypothesis: Learning by contiguity, responding by relative strength'. In: *Handbook of Contemporary Learning Theories*, R.R. Mowrer & S.B. Klein (eds), pp. 65–117. Mahwah, NJ: Erlbaum.

Dickinson, A. (2001) 'Causal learning: An associative analysis', *The Quarterly Journal of Experimental Psychology* 54B: 3–25.

—— & Burke, J. (1996) 'Within-compound associations mediate the retrospective revaluation of causality judgements', *Quarterly Journal of Experimental Psychology* 49B: 60–80.

Frosch C.A., McCormack, T., Lagnado, D.A., & Burns, P. (in press) 'Are causal structure and intervention judgments inextricably linked?' *Cognitive Science*.

[4] The research reported in this chapter was supported by grants from the Economic and Social Research Council and the Arts and Humanities Research Council, UK.

Gopnik, A., Glymour, C., Sobel, D.M., Schulz, L.E., Kushnir, T., & Danks, D. (2004) 'A theory of causal learning in children: causal maps and Bayes nets', *Psychological Review* 111: 3–32.

—— & Schulz, L.E. (2007) (eds) *Causal Learning: Psychology, Philosophy, and Computation*. Oxford: Oxford University Press.

—— Sobel, D.M., Schulz, L.E., & Glymour, C. (2001) 'Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation', *Developmental Psychology* 37: 620–29.

Hagmayer, Y., Sloman, S., Lagnado, D., & Waldmann, M.R. (2007) 'Causal reasoning through intervention'. In: *Causal Learning: Psychology, Philosophy, and Computation*, A. Gopnik & L.E. Schulz (eds), pp. 86–100. Oxford: Oxford University Press.

Harris, P.L., German, T., & Mills, P. (1996) 'Children's use of counterfactual thinking in causal reasoning', *Cognition* 61: 233–59.

Hart, H.L.A. & Honoré, A.M. (1985). *Causation in the Law*. Oxford: Clarendon Press.

Kruschke, J.K. (2006) 'Local Bayesian learning with applications to retrospective reevaluation and highlighting', *Psychological Review* 113: 677–99.

—— & Blair, N.J. (2000) 'Blocking and backward blocking involve learned inattention', *Psychonomic Bulletin and Review* 7: 636–45.

Kushnir, T. & Gopnik, A. (2007) 'Conditional probability versus spatial contiguity in causal learning: preschoolers use new contingency evidence to overcome prior spatial assumptions', *Developmental Psychology* 43: 186–96.

Lagnado, D.A. & Sloman, S.A. (2004) 'The advantages of timely intervention', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30: 856–76.

Le Pelley, M.E., Oakeshott, S.M., & McLaren, I.P.L. (2005) 'Blocking and unblocking in human causal learning', *Journal of Experimental Psychology: Animal Behavior Processes* 31: 56–70.

Lewis, D. (1973) 'Causation', *Journal of Philosophy* 70: 556–67.

Lovibond, P.F., Been, S., Mitchell, C.J., Bouton, M.E., & Frohardt, R. (2003) 'Forward and backward blocking of causal judgment is enhanced by additivity of effect magnitude', *Memory & Cognition* 31: 133–42.

McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009) 'Cue competition effects and young children's causal and counterfactual reasoning', *Developmental Psychology* 45: 1563–75.

Mackie, J.L. (1974) *The cement of the universe*. Oxford: Oxford University Press.

Mitchell, C.J., Lovibond, P.F., & Condoleon, M. (2005) 'Evidence for deductive reasoning in blocking of causal judgements', *Learning and Motivation* 36: 77–87.

Perner, J., Sprung, M., & Steinkogler, B. (2004) 'Counterfactual conditionals and false belief: a developmental dissociation', *Cognitive Development* 19: 179–201.

Pineño, O. & Miller, R.R. (2007) 'Comparing associative, statistical, and inferential reasoning accounts of human contingency learning', *Quarterly Journal of Experimental Psychology* 60: 310–29.

Riggs, K.J., Peterson, D.M., Robinson, E.J., & Mitchell, P. (1998) 'Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality?', *Cognitive Development* 13: 73–90.

Roser, M.E., Fugelsang, J.A., Dunbar, K., Corballis, P.M., & Gazzaniga, M.S. (2005) 'Dissociating processes supporting causal perception and causal inference in the brain', *Neuropsychology* 19: 591–602.

Schlottmann, A. (1999) 'Seeing how it happened and knowing how it works: How children understand the relationship between perceptual causality and underlying mechanism', *Developmental Psychology* 35: 303–17.

Schlottmann, A. (2000) 'Is perception of causality modular?' *Trends in Cognitive Sciences* 4: 441–42.

—— & Shanks, D.R. (1992) 'Evidence for a distinction between judged and perceived causality', *Quarterly Journal of Experimental Psychology,* 44A: 321–42.

Schulz, L.E., Gopnik, A., & Glymour, C. (2007) 'Preschool children learn about causal structure from conditional interventions', *Developmental Science* 10: 322–32.

—— Kushnir, T., & Gopnik, A. (2007) 'Learning from doing: Intervention and causal inference'. In: *Causal Learning: Psychology, Philosophy, and Computation,* A. Gopnik & L.E. Schulz (eds), pp. 67–85. Oxford: Oxford University Press.

Sedlak, A.J. & Kurtz, S.T. (1981) 'A review of children's use of causal inference principles', *Child Development* 52: 759–84.

Shanks, D.R. (1985) 'Forward and backward blocking in human contingency judgment', *Quarterly Journal of Experimental Psychology* 37B: 1–21.

—— (2007) 'Associationism and cognition: Human contingency learning at 25', *Quarterly Journal of Experimental Psychology* 60: 291–309.

Shultz, T.R. (1982) *Rules for causal attribution. Monographs of the Society for Research in Child Development* 47 (1, Serial No. 194).

Sloman, S.A. (2005) *Causal Models: How People Think about the World and its Alternatives.* Oxford: Oxford University Press.

—— & Lagnado, D.A. (2005) 'Do we "do"?', *Cognitive Science* 29: 5–39.

Sobel, D.M. & Kirkham, N.Z. (2006) 'Blickets and babies: The development of causal reasoning in toddlers and infants', *Developmental Psychology* 42: 1103–15.

—— Tenenbaum, J.B., & Gopnik, A. (2004) 'Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers', *Cognitive Science* 28: 303–33.

Steyvers, M., Tenenbaum, J.B., Wagenmakers, E.J., & Blum, B. (2003) 'Inferring causal networks from observations and interventions', *Cognitive Science* 27: 453–89.

Vandorpe, S., De Houwer, J., & Beckers, T. (2005) 'Further evidence for the role of inferential reasoning in forward blocking', *Memory & Cognition* 33: 1047–56.

—— —— —— (2007) 'Outcome maximality and additivity training also influence cue competition in causal learning when learning involves many cues and events', *Quarterly Journal of Experimental Psychology* 60: 356–68.

Wasserman, E.A. & Berglan, L.R. (1998) 'Backward blocking and recovery from overshadowing in human causal judgment: the role of within-compound associations', *Quarterly Journal of Experimental Psychology* 51B: 121–38.

White, P.A. (1995) *The Understanding of Causation and the Production of Action: From Infancy to Adulthood.* Hove, East Sussex: Erlbaum.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation.* Oxford: Oxford University Press.

—— (2007) 'Interventionist theories of causation in psychological perspective'. In: *Causal Learning: Psychology, Philosophy, and Computation,* A. Gopnik & L.E. Schulz (eds), pp. 19–36. Oxford: Oxford University Press.

3

Perceptual Causality, Counterfactuals, and Special Causal Concepts

Johannes Roessler

On one view, an adequate account of causal understanding may focus exclusively on what is involved in mastering general causal concepts (concepts such as '*x* causes *y*' or '*p* causally explains *q*'). An alternative view is that causal understanding is, partly but irreducibly, a matter of grasping what Anscombe called special causal concepts, concepts such as 'push', 'flatten', or 'knock over'. We can label these views generalist vs particularist approaches to causal understanding. It is worth emphasizing that the contrast here is not between two kinds of theories of the metaphysics of causation, but two views of the nature and perhaps source of ordinary causal understanding. One aim of this paper is to argue that it would be a mistake to dismiss particularism because of its putative metaphysical commitments. I begin by formulating an intuitively attractive version of particularism due to P.F. Strawson, a central element of which is what I will call naïve realism concerning mechanical transactions. I will then present the account with two challenges. Both challenges reflect the worry that Strawson's particularism may be unable to acknowledge the intimate connection between causation and counterfactuals, as articulated by the interventionist approach to causation. My project will be to allay these concerns, or at least to explore how this might be done. My (tentative) conclusion will be that Strawson's naïve realism can accept what interventionism has to say about ordinary causal understanding, and that interventionism should not be seen as being committed to generalism.

1 Special Causal Concepts and Perception

Suppose you observe a boulder rolling down the mountainside and flattening the wooden hut in its path (Strawson, 1992: 118) We can ask two questions about the causal judgement you are likely to make as a result of this experience. What is the content of your judgement? And what is its basis or source?

The standard answer to the first question is that your judgement will be of the form '*x* causes *y*' and will involve reference to two observed events. It is in these terms that 'perceptual causality' tends to be characterized both in the psychological and in the philosophical literature. For example, according to Schlottmann and Shanks, in a Michottean launch event 'successive motion of two objects appears to evoke an immediate perception that the first motion caused the second' (1992: 321). Turning to the second question, the familiar problem here is what to make of the idea that it is 'immediate perception' that gives rise to the judgement. One reason why the idea can seem hard to take very seriously is this: to say that the relation between the two events is causal is to say that it is not a mere correlation. It is not just, for example, that the two events occur as the result of a common cause. Let us call the question whether an observed sequence of events is related as cause and effect, or whether it is a mere correlation, the *causal question*. Now what should count as an adequate basis for answering the causal question arguably depends on what the difference between causes and correlations amounts to or consists in. An attractive approach to this latter question is that the difference has to do with the correctness of certain counterfactual conditionals. As Hume said, 'if the first object had not been, the second never had existed' (Hume, 1975: 76). The idea has recently been given an authoritative formulation by the interventionist analysis of causation. Very briefly, the suggestion is that the difference between causal relations and mere correlations turns on the correctness of suitable 'intervention counterfactuals', to the effect, roughly, that had there been an intervention on the value of the cause variable this would have made a difference to the value of the effect variable. Given this analysis, the problem then is that it is difficult to see how you can extract information about the correctness of such counterfactuals from observation of a sequence of events in a single case, unless of course perception simply supplies some relevant input to your theorizing about the situation. In that case, though, it would be misleading to characterize the basis of your judgement as 'immediate perception'.

Strawson rejects the standard answer to the first question. In its place he recommends the following alternative: you will judge that *the boulder flattened the hut*. Does this alternative answer to the first question (concerning the content of your judgement) make a material difference to the way we should think about the second question (concerning its basis)? Strawson certainly thinks it does. In this section I want to set out in some detail his position, its commitments, and some of the concerns that motivate it. In the next two sections, I will further probe his claim about the bearing of his account on the second question.

Strawson's account of the content of our judgement differs from the standard picture in two respects. First, on his account, your judgement involves not the general concept of causation but what Anscombe calls a special causal concept. Her famous list of such concepts includes the following: 'scrape, push, wet, carry, eat, burn, knock over, keep off, squash, make (e.g. noises, paper boats), hurt' (Anscombe, 1993: 93). Strawson characterizes such concepts as concepts of 'kinds of action and transaction' (1992: 116), of 'varieties of bringing something about, of producing some effect or some new state

of affairs' (1992: 115). The second difference concerns the ontology of causation. While the standard picture takes your judgement to be about a causal relation between *two events* (the boulder colliding with the hut and the hut collapsing), on Strawson's reading, the judgement is about a causal transaction between two *physical objects* (the boulder and the hut), where the causal transaction constitutes a *single event* (the boulder flattening the hut) that causally explains a certain fact or state of affairs (that the hut has been left flat). We might call the two positions on the first question the two events view and the single event view, respectively.

As we saw, on the two events view, the perplexing feature of 'perceptual causality' is that 'immediate perception' *appears* to yield an answer to the causal question, the question of whether the relation between two given events is causal or merely correlational. Strawson maintains that this way of thinking about the matter is misconceived: 'There is no question of dissolving the transaction into a sequence of states of affairs—a sequence of "distinct existences"—and wondering whether, or in virtue of what, the sequence constitutes a causal sequence. One has observed the change being brought about in some characteristic mode' (1992: 116). The idea here is that a judgement about a causal transaction between two physical objects can intelligibly arise from 'immediate perception', given that in perception we may be aware of the (single) event of the two objects interacting in a specific mode. The shift to special causal concepts and the single event view is not intended to explain how perception can provide the answer to the causal question. The idea, rather, is that the causal question *does not normally arise* when we perceive mechanical transactions among physical objects, and does not arise *because* we conceptualize the situation in terms of special causal concepts.[1]

[1] It is important to distinguish Strawson's view from certain theories of causation that give a privileged role to mechanical processes. In an illuminating recent paper, James Woodward contrasts 'geometrical/mechanical' theories of causation with 'difference-making' theories. The former focus 'on cases in which there is a physical process connecting cause and effect and, more broadly, on phenomena that are mediated by contact-mechanical forces and in which spatio-temproal or geometrical relationships play an important role' (2011: p. 203f). The latter appeal to 'difference-making information', such as information on what happens to the effect under interventions on its putative cause. According to Woodward, the appeal of geometrical/mechanical theories, such as it is, is grounded, at least in part, in intuitions concerning 'perceptual causality': they promise to make the possibility of perceptual knowledge of causal relations less perplexing than it would be under a difference-making account. Now there are two reasons it would be mistaken to group Strawson's particularism with geometrical/mechanical theories of causation in Woodward's sense. First, Strawson is not primarily concerned with the metaphysical issue of the nature of causation (which is what both of the two kinds of theories Woodward distinguishes are about) but with the source of ordinary causal understanding. Second, on Strawson's view, we cannot get the source of causal understanding right if we focus exclusively on the general concept of causation. His emphasis on mechanical transactions is an emphasis on a particular class of *special* causal concepts.

Admittedly, Strawson suggests that the fundamental role of such concepts in our causal thinking may account for certain psychological facts about us; for example, our reluctance to countenance action at a distance, or our tendency to seek to assimilate causal phenomena in other areas to the case of mechanical transactions—intuitions that might be invoked in support of geometrical/mechanical theories of causation. But I think this part of Strawson's discussion is intended to be an exercise in psychology, not metaphysics (and it is anyhow something of an afterthought). It is not a commitment of Strawson's view that causation in general has to be conceived on the model of mechanical transactions.

One ingredient of the view Strawson is recommending here might be called naïve realism about mechanical transactions. This involves both phenomenological and epistemological commitments, and indeed commitments concerning the relation between experience and knowledge. Thus a naïve realist may claim that mechanical transactions are visible in the sense that we have visual recognitional capacities for them. This does not require that such transactions are amongst the properties and relations of things that help to *constitute* the character of our visual experience of them, alongside shape, colour, orientation, motion, and so on. What it arguably does require is that our ability visually to recognize mechanical transactions is immediately *intelligible* to us in terms of the character of our visual experience, constituted by those lower-level properties. For example, we find your ability to tell that the boulder flattened the hut unmysterious in the light of the way things looked to you when you observed the transaction. (See Roessler, 2011 for a more detailed discussion of the relation between experience and recognitional capacities.) This picture goes beyond anything Strawson explicitly says; but I think it is in the spirit of his position.

Importantly, Strawson combines naïve realism concerning mechanical transactions with the rejection of naïve realism concerning causal relations between distinct events. Suppose you have become adept at switching back and forth between two ways of experiencing a Michottean launch phenomenon: you are able to see it now *as one event* (one object pushing away another object) and now *as a sequence of two events* (a collision between two objects followed by movement on the part of the second object). Then, on Strawson's view, in switching from the first to the second mode of experience there is a sense in which you disable yourself from perceiving causation. In seeing (as such) the single event of one object acting on another, you see a specific kind of causal transaction. In seeing the display as a sequence of two events you merely see two spatially and temporally related events: your perception will raise, but will not on its own be able to answer, the question of whether the sequence is a causal sequence. Now one of Strawson's claims is that this latter kind of experience is uncommon. To suggest otherwise, as the two event view does, is to 'atomize' and thereby 'falsify' the phenomenology of visual and tactual experience of mechanical transactions (1992: 118). The following analogy may help to clarify Strawson's thinking here. Elsewhere he argues that ordinary visual experience presents itself as an immediate experience of physical objects, objects that do not depend for their existence on being perceived. But he acknowledges the possibility of a radical Gestalt switch: 'An observer, gazing through his window, may perhaps, by an effort of will, bring himself to see ... what he knows to be the branches of the tree no longer *as* branches at all, but as an intricate pattern of dark lines of complex directions and shapes and various sizes against a background of varying shades of grey' (1988: 90). The point of this concession is to bring out the contrast between this contrived mode of experience and ordinary perceptual consciousness, an adequate account of the character of which has to make use of concepts of physical objects. The dialectical role of this point is to help expose what Strawson regards as a fundamental distortion. The distortion consists in

misrepresenting our commitment to the existence of physical objects as a *theoretical* commitment, based on evidence supplied by perceptual experience. What this view assumes is that there can be a completely general question as to how the commitment is to be justified—and how the concepts that go into it may be acquired—on the basis of the more primitive deliverances of perceptual experience. However, given the actual character of our (ordinary, non-contrived) perceptual consciousness, such a question cannot arise: a commitment to the existence of physical objects is 'given with the given' (1988: 99). The analogous claim about causation is that uncompromising opponents of naïve realism misrepresent both the character of our ordinary experience of mechanical transactions and—connectedly—the status of commonsense realism about causation. They 'atomize' the way we experience mechanical transactions, and they regard our commitment to the reality of causation as the result of implicit or explicit reasoning about the causal question, relying on evidence and inference, rather than as something that is 'given with the given'.

2 Interventionism

The task of assessing these arguments for naïve realism concerning mechanical transactions and physical objects, respectively, is well beyond the scope of this paper. (But I will come back to the relation between them in the next section.) In any case, I suspect the immediate concern prompted by Strawson's version of naïve realism is not so much that it has no attractions as that it may be open to decisive objections. On Strawson's view, our understanding of causation is embodied, partly but irreducibly, in our grasp of special causal concepts. This is what I called a particularist view of causal understanding. Those who are sceptical about this view do not deny that a great deal of causal thinking is done by the use of special causal concepts. But they see no explanatory value in focusing on such concepts when considering the question of what is involved in understanding causation. On the generalist view, an answer to that question has to draw on an analysis of the general nature of causation. Questions about ordinary causal understanding should be understood as questions about the way in which the metaphysical facts of causation are available to creatures capable of causal thinking.

Suppose that in filling in the generalist picture we take the interventionist analysis of the metaphysics of causation as our starting point. It seems to me that this would be an excellent choice, though there is no space to defend that choice or even to provide a detailed exposition of interventionism. (One rather quick way to motivate the choice would be to note how influential interventionism has been in recent work on causal understanding in both philosophy and psychology.) I take the core idea of an interventionist approach to causal understanding to be this:

INT-CU: To understand what it is for X to cause Y is to appreciate that were there to be an intervention on the value of X there would be a change in the value of Y.

Put loosely and intuitively, an intervention on X is a causal relation between some variable I and X such that I causes a change in the value of X in an 'exogenous' way—it takes complete control of X without affecting Y in any other way than through X. Leaving aside the critical issue of how to make that intuitive notion more precise (see Woodward (2003) for a sustained discussion), the question that matters in the current context is what is involved in 'appreciating' the entailment between causal claims and interventionist counterfactuals. It would be quite unsatisfactory to construe this as a matter of explicit propositional knowledge, if only because the notion of an intervention is a technical philosophical notion that can hardly be assumed to be familiar to all. A much more plausible suggestion is this. An interventionist account of causation seeks to illuminate the content of causal claims by articulating in a perspicuous manner the commitments we incur in making such claims. One shows one's 'appreciation' of such commitments through the patterns of reasoning one accepts as relevant when establishing causal claims and when drawing out implications of causal claims. This is not the only way in which one might bring interventionism to bear on ordinary causal understanding. But it seems to me the most plausible model, and I think it is the one Woodward himself favours.[2] It is worth stressing that this model offers no reductive account of causal understanding, any more than interventionism offers a reductive account of the nature of causation. In both cases, causal concepts figure conspicuously in the analysis. For example, exploiting the affinity between the technical notion of an intervention and the commonsensical concept of manipulation, an interventionist is likely to argue that a basic way to manifest one's possession of causal concepts is by appreciating the *practical* implications of causal claims, by acting intentionally on causes in order to manipulate their effects. In effect, what we are talking about here is the idea that causal understanding involves being able to appreciate certain *causal* implications of causal claims. It is by no means clear, though, that this should be seen as an objection to the commitment model. Intuitively, the model is neither trivial nor unilluminating. It is not obvious that a reductive account of causal understanding is compulsory.

We can distinguish two ways in which the interventionist account may put pressure on Strawson's particularism. According to what I'll call the manipulation challenge, there is a completely general connection between grasp of causal concepts (whether general or special causal concepts) and an appreciation of the implications of causal claims for purposes of manipulation. The particularist account, it may be said, fails to respect that connection. It should be rejected as an 'impractical' theory of causal understanding.[3] According to what I will call the justification challenge, there is a mismatch between the *basis*, as construed by the particularist, and the *content* of causal claims. Even an ostensibly perceptual judgement about a particular boulder's flattening a particular hut commits us to general counterfactual claims to do with certain features,

[2] See especially Woodward (2007).

[3] See Woodward (2003: 30) for objections, under that heading, to a range of theories of causation (though not Strawson's).

or the general types, of the objects involved. For causation implies a *reproducible* relationship between variables. It requires that changing the value of one variable is a way of manipulating the value of the other variable in certain kinds of circumstances. Moreover, it is in virtue of our understanding of some such type-causal claim that we are entitled to make counterfactual judgements about a particular case (e.g. that 'if the first object had not been, the second never had existed', or perhaps, would still exist). Given the counterfactual commitments inherent in the content of causal claims, it is wholly implausible that passive observation of a single causal transaction could provide an adequate basis for such claims.

I briefly mention an extreme line of response, though only to set it to one side. One might seek to make particularism immune to these challenges, by insulating special causal concepts and 'actual causation' from the sorts of commitments affirmed by the critic. Anscombe writes that 'the word "cause" can be *added* to a language in which are already represented many causal concepts' (1993: 93). This may be read—though perhaps the passage is open to more than one interpretation—as suggesting that special causal concepts are autonomous in this sense: their mastery is independent of, and prior to, grasp of causation in general. This, in turn, might suggest that such concepts are available independently of appreciating any general connection between causal claims and manipulation, contrary to the manipulation challenge. A similarly extreme response to the justification challenge would be to adopt a singularist approach to causation (also often associated with Anscombe), on which singular causal claims do not commit us to any type-causal claims at all.

I want to suggest, or at least explore the possibility, that particularism may be defensible without resorting to such heroic measures. In the next section, I discuss the manipulation challenge. In Section 4 I return to the—arguably more serious—justification challenge.

3 Causal Understanding and Manipulation

I take it to be common ground that claims couched in special causal terms have *a priori* determinable causal implications.[4] That the boulder flattens the hut implies that the boulder brings about, or causes, a certain outcome or effect. Presumably this is the rationale for calling the concept 'flatten' a special *causal* concept in the first place. It is a further question whether possession of the concept requires knowledge, explicit or in some sense implicit, that the implication holds. I want to suggest that it is open to a particularist to accede to this requirement. To do so is not to accept the idea that mastery of a given special causal concept can be reduced to understanding causation in general, plus suitable non-causal material (grasp of some relevant spatial and temporal concepts, say). An alternative picture might be that the relation between understanding

[4] I borrow this formulation from Paul Snowdon (2011).

special and general causal concepts is one of *mutual dependence*. One doesn't fully grasp what it means for one object to push or scrape or flatten another unless one recognizes the manipulationist implications of such interactions, showing one's appreciation of certain completely general commitments of causal claims.[5] Still, one's understanding may have its source, in part, in one's observation of (and perhaps participation in) specific modes of interaction among physical objects.[6]

I want to bring out one way in which this reciprocal view bears on the development of causal understanding. Consider the question whether infants' engagement with Michotte-type launching phenomena, as probed in looking-time studies, should be interpreted as a manifestation of causal beliefs, i.e. possession of causal concepts. Leslie and Keeble (1987) hint at an affirmative answer. On their interpretation, their experimental work provides evidence that 'causal understanding' has 'its beginnings in a low level visual mechanism' (1987: 285). An interventionist response to this claim would be that the relation of what is probed in looking-time studies to propositional attitude psychology is no straightforward matter; and that manifestations of even the beginnings of causal understanding should be expected to include an appreciation of the implications of causal claims for manipulative activities. (See Woodward, 2011.) From the interventionist perspective, the development of causal understanding is inextricably bound up with the development of counterfactual reasoning, a minimal form of which may simply involve the use of future hypotheticals in practical reasoning. (See the papers by Woodward; Beck et al.; McCormack et al.; Perner and Rafetseder in this volume.)

The reciprocal view concurs with this line of response, but points out that this is entirely consistent with *rejecting* a second kind of interventionist comment on 'perceptual causality'. The second comment claims that counterfactual reasoning is required not just to grasp the essential *commitments* of causal claims, but to marshal the canonical *evidence* on which to base such claims. This second comment assumes that perception alone can provide no adequate basis to determine whether the relationship between two observed events is merely correlational or causal. It argues that the canonical method to settle this is to design and carry out a suitable experiment, with a view to obtaining evidence for the relevant interventionist counterfactuals. Put bluntly, on this view the case of observing mechanical interactions is not in principle different from the diffuse 'impression of causality' one might enjoy when opening the door at the same moment the light comes on. (See Schlottmann, 2001.) The difference is merely a psychological, not an epistemological matter. As Woodward writes, in certain cases

[5] One reason particularists may be well-advised to resist Anscombe's autonomy claim may be the thought that understanding special causal concepts requires the ability to give causal explanations, and that the latter in turn requires some grasp of the completely general idea of one thing causally explaining another. I suspect Strawson would be sympathetic to this thought. He certainly emphasizes that directly observed causal transactions provide adequate causal explanations of their outcome.

[6] The relation between the respective contributions of observation and participation in mechanical interactions as sources of causal understanding is an issue that would deserve more extensive discussion. Strawson focuses on the former, but also stresses the importance of the latter and of possible interconnections between the two.

'the possibility that the relationship in question [is] merely correlational seems so unlikely that it is easy to overlook the point that a separate inferential step is required to move from the claim that two kinds of events are correlated to a claim about how one will respond under manipulation of the other' (2003: 33). The reciprocal view maintains that this is to misrepresent the epistemic role of perceptual experience. Observing mechanical transactions does not present us with the question of how to understand the relationship between two events. For it does not present us with two events at all, but with the single event of an object pushing or flattening or squashing another object. Thus it may enable us to perceive without reliance on inference and experimental evidence *that* one object is pushing or flattening or squashing another. The development of causal understanding may be a matter, partly, of acquiring (or actualizing) the ability to perceive mechanical transactions as such.[7] This is of course not to deny that an equally important aspect of that development is the ability to evaluate causal claims concerning less immediately accessible matters on the basis of experimental evidence.

A defender of the second sort of interventionist comment on 'perceptual causality' may well ask why we *should* treat 'causal impressions' of pushings and flattenings differently from perceptually induced hunches. Furthermore, even granting that the cases appear to be different, he may remind us of the justification challenge: how can perceptual experience on a single occasion provide an adequate justification for the characteristic counterfactual commitments of causal claims?

4 Dispositions and Their Grounds

Recall Strawson's remark that in observing mechanical transactions, 'there is no question of dissolving the transaction into a sequence of states of affairs—a sequence of "distinct existences"—and wondering whether, or in virtue of what, the sequence constitutes a causal sequence. One has observed the change being brought about in some characteristic mode' (1992: 116). The fact remains, it might be said, that even if an observer does not address the causal question, even implicitly, but instead describes the situation by the use of special causal concepts, her judgement *commits* her to a certain answer to the causal question. The judgement 'the boulder flattened the hut' commits one to the obtaining of a causal relation between two events, the observed collision and the attendant collapse of the hut. And to claim that the relationship between these events is causal rather than merely correlational is to be committed to the correctness of counterfactuals concerning *reproducible* relations between interventions on the first *type* of event and changes in the second *type* of event. If the observer's causal judgement is rich in these sorts of general commitments, how can observation of a causal transaction on a single occasion provide an adequate basis for it?

[7] I think the point is neutral on debates in psychology over the existence and role of innate constraints on the development of causal understanding.

Although Strawson does not explicitly address this challenge, his discussion contains two moves that are highly relevant to it. The first move is to draw attention to the fact that in perceiving mechanical transactions we also tend to perceive some of the *features* of objects in virtue of which they have the causal powers exercised in such transactions. The second move consists in the claim that our observational concepts of such features are *dispositional* concepts: it is 'internal' to such concepts that 'the bearers of those qualities regularly act and react in such-and-such ways' (1992: 122). Strawson presents this latter point as a partial concession to (e.g. Humean) opponents of naïve realism concerning mechanical transactions: 'With this dispositionality, the *generality*, which is the core of the reductive conception, is already given' (ibid.). But the concession is not expected to supply 'any comfort to the reductionist':

> This it could do only if the relevant concepts of thing and quality dissolved into, or were constructed out of, a complex of wholly non-dispositional concepts (of sense-quality) plus generalizations relating them. But the relevant concepts of thing and quality do not so dissolve and are not so constructed. They are basic conceptual stock; and to think otherwise is to misrepresent us as theorists before we have the means of theorizing (1992: 122).

I want to make three brief comments on the first move, before raising worries about the second move and sketching an alternative to it.

Here is a classical example of a mechanical interaction. Suppose you see a square axle passing through a square (and similarly sized) hole in a millstone. And suppose you then see the axle turn. According to the naïve realist, in seeing the axle turn you will see the axle *turning the millstone*. Your experience provides you with an immediate explanation of the fact that the millstone is moving. (The axle is turning it.) Strawson's first move may be put by saying that the availability of this explanation is essentially connected with that of another sort of explanation. You perceive certain features of both the millstone and the axle that make their interaction *intelligible*: most importantly, their respective shapes and their solidity. For it is in virtue of these properties that the axle has the power to turn the millstone.

My first comment is that *if* our perceptual experience supplied us with both sorts of explanation, this would indeed help to make it unmysterious how perception can be a source of knowledge of the kinds of counterfactuals implicit in causal claims. If we could single out some of the features in virtue of which the axle has the power to move the millstone this would give us at least a rough idea of what sorts of interventions on the axle would make a difference to the movement of the millstone. None of this is to suggest that your perceptually based judgement that the axle turned the millstone could not be mistaken. The point of drawing attention to the intelligibility of mechanical interactions in terms of perceived features is not to rule out the possibility of error and illusion—something that is undeniable but is of little importance in the current context. For the possibility of illusions of causal transactions does not prove that in a veridical case we are presented with something less than particular modes of causal transaction. The reason it matters that perceived transactions are intelligible in terms of

perceived features is that the perception of the relevant features may ground the kind of knowledge that opponents of naïve realism claim has to rely on evidence (for example, experimental evidence) and theorizing.

Second, the point surely holds the key to the question of what is special about mechanical transactions. The obvious difference between 'perceptual causality' in the mechanical case and the sorts of hunches one might have when opening a door at the precise moment the light comes on is that in the latter case one's perceptual experience is silent on the features in virtue of which the door might have the power to activate the light. Part of what makes such 'causal impressions' mere hunches is that the relevant causal relation is not intelligible in terms of perceived properties. The difference between the two sorts of cases is an epistemological matter, not merely a matter of the 'robustness' of the phenomenology of the causal impression.

Third, earlier I discussed a parallel or analogy between Strawson's case for naïve realism concerning mechanical transactions and his case for naïve realism concerning physical objects. In both instances, opponents of naïve realism are said to distort the character of ordinary perceptual experience and, connectedly, to misrepresent as a theoretical commitment something that is 'given with the given'. There are some indications that Strawson himself takes the two cases to be rather more closely connected—that he takes naïve realism about mechanical transactions to be *entailed* by naïve realism about physical objects.[8] I cannot pursue this idea, but one way one might try to reach such a conclusion would be to argue thus: physical objects can be immediately perceived only in so far as some of their properties, including shape and solidity, are immediately perceivable also; and the satisfaction of this latter condition makes it possible for mechanical transactions amongst physical objects to be immediately perceivable too.

Now to meet the justification challenge, it is not enough to point out that we directly perceive features such as shape and solidity, and that appeal to such features may help to make mechanical transactions intelligible. An immediate line of response is that even if perceptual experience reveals to us what shapes are and thus enables us to think about them, it takes experiment and theory to work out what sorts of dispositions are associated with a given shape. Strawson's second move seems designed to counter this response. It claims that our concepts of features such as shape or solidity are *both* observational *and* dispositional. This claim is not easy to understand, though. I want to examine briefly what I take to be the most promising line of defence, though only to conclude that Strawson's view cannot be sustained. I end with a tentative suggestion as to how one might tackle the justification challenge without reliance on Strawson's peculiar view of observational concepts.

[8] Compare his remark that Kant 'had a secure grasp of the central point, which it would perhaps not be unacceptable to express in more or less his own words by saying that the concept of causal efficacy is not derived from experience of a world of objects, but is a presupposition of it; or, perhaps better, is already with us when anything which could be called "experience" begins' (1992: 124).

For simplicity I focus on the case of shape. An obvious worry about Strawson's second move is that ordinary shape concepts do not appear to be dispositional concepts. They are not concepts of tendencies of objects to respond in certain ways to certain circumstances. They are not like concepts such as brittle or bendable. To develop and defend Strawson's position here one might invoke a view of properties advanced by Sydney Shoemaker. On this view, properties are what Shoemaker calls clusters of conditional powers. Shoemaker acknowledges that shapes, for example, are not dispositions in the ordinary sense. But he argues that what determines the identity of a shape property is the 'potential for contributing to the causal powers of the things that have it' (2003: 212). For example, to be knife-shaped is, amongst other things, to have the power of cutting butter and cheese, conditionally upon its having certain other properties, such as being made of steel and being knife-sized. He labels the powers an object has conditionally on having certain other properties its conditional powers. Shoemaker's main reason for the claim that properties, even shape properties, are conditional powers is that he regards this as the only way of avoiding the unpalatable consequence of making properties wholly inaccessible to us. If shape properties were logically distinct from the cluster of conditional powers things have in virtue of having a particular shape, we would be cut off from their presence. It would be unintelligible 'how properties are capable of engaging our knowledge, and our language, in the way they do' (2003: 214).

In a crucial respect, Shoemaker's view is clearly congenial to Strawson. On Shoemaker's view, it is impossible to analyse our conception of shape properties into grasp of non-dispositional concepts of 'sense-quality' plus empirical generalizations as to how objects having such qualities tend to act and react under particular sorts of circumstances. You have no grasp of a property unless you have mastered something like a functional characterization of the property. This chimes with Strawson's remarks about the 'thoroughgoing dispositionality' of our concepts of things and their qualities (1992: 122).

The trouble is that there are other aspects of Shoemaker's view that are in tension with Strawson's naïve realism. How do we come to know what shape properties are? For Shoemaker this is a matter, most fundamentally, of observing 'actualizations' of the causal potentialities definitive of a given property. 'We know and recognize properties by their effects, or, more precisely, by the effects of the events which are the activations of the causal powers which things have in virtue of having the properties' (2003: 214). Shoemaker offers a subtle account of how knowledge of even a relatively small subset of the conditional powers definitive of a property can be sufficient to 'fix the reference' of a property term. Without going into the details, the basic picture is that to grasp the property of being square you need to understand some patterns of causal relations, such as the causal relations between (a) the rotation of a square axle passing through a square hole in the millstone and (b) the rotation of the millstone; or the 'causal connection between the possession of [the property of being square] by an object and the sensory states of an observer related to that object in certain ways, e.g. looking at it in good light' (2003: 214).

There are several reasons why Strawson has to reject this picture. Shoemaker's view might be described as a 'theory theory' of our grasp of shape properties. We come to know what it is for something to be square by discovering something about the functional role of the property of being square. This picture implies that in acquiring shape properties we emerge from just the kind of condition Strawson declares to be impossible: a stage at which we are aware of objects and events but are 'equally prepared, or unprepared, for anything to come of anything' (1992: 122). The twist Shoemaker gives to this view is that what is lacking at this primitive stage is not only knowledge of the causal powers of things but also, connectedly, mastery of shape concepts. Importantly, though, one thing he thinks is available even at this stage are concepts of shape *appearances*: acquiring concepts of the shapes of physical objects is partly a matter of understanding what kinds of sensory appearance they tend to cause in suitably related observers. This is another reason Strawson could not accept the view of properties as clusters of conditional powers. An accurate description of the character of sensory appearances, according to Strawson, requires *using* concepts of some of the properties of perceived physical objects. (See Strawson, 1988.) I conclude that an attempt to secure a solid metaphysical footing for Strawson's view of shape concepts is unlikely to be successful.

As John Campbell has pointed out, we ordinarily think of the shape of an object not as a cluster of dispositions to act or react in certain ways but as the causal ground of those dispositions (2002: ch. 12). We can put the point in interventionist terms. Suppose you intervene on the shape of Arnauld's axle. Suppose, say, you change the value of the variable 'shape' from 'square' to 'round'. By doing so you will in one fell swoop manipulate an extremely large variety of dispositions. For example, you would change the axle's disposition to turn a wide range of objects, conditional on being suitably installed. Or, moving on to another classic example, you would change the axle's disposition not to pass through a round hole (where the length of its side is the same as the diameter of the hole) (Putnam, 1975). Intuitively, the axle's shape provides the causal explanation of its disposition to act and react in these myriad specific modes. Suppose we accept this intuitive view of shape, rather than Shoemaker's revisionist account. Does it then follow that contra Strawson, grasp of what it is for something to be square is not sufficient for knowledge of any of the dispositions square objects have to act and react in specific ways? Does it follow that understanding shape concepts is consistent with being 'equally prepared, or unprepared, for anything to come of anything'—a position from which one can only emerge by constructing or assimilating an empirical theory of the causal powers of shapes, based on repeated observation or experimental evidence?

It is not obvious that the intuitive view has this implication. There is a time-honoured view according to which the explanatory connection between the shape of an object and its disposition for mechanical interactions with other objects (conditional on its other properties) is not an empirical matter. It is this idea that Arnauld's example was originally intended to illustrate. In Michael Ayers' words, 'whether keys

will turn locks or axles will turn mill-stones can simply be evident from the shape of the rigid parts, together with the idea of a body's defining capacity to push others out of its way' (1991: 147). Arnauld unhelpfully put the point in terms of infallibility ('l'effet qu'il a prétendu s'ensuit infailliblement'[9]); a formulation echoed in Hume's discussion of the idea of a power. But of course in any particular case the effect may fail to occur for some undetected reason. The 'evidence' or 'perspicuity' illustrated by the example really concerns a *type*-causal claim: it can be evident *a priori* that the property of being square causally explains the disposition to turn a (suitably installed) millstone. It is often assumed that Hume's treatment of causation has rendered this rationalist view wholly obsolete. But as Ayers points out, it is at least a good question why 'mechanical interactions appear more intelligible to us than other experienced physical changes'. The Humean answer, 'that this is simply a matter of familiarity', is 'utterly unconvincing': 'We are not much less used to the effects of heat and light than to the effects of the mechanical properties of things. Yet we do not immediately understand why heat will harden an egg as we can immediately understand why Arnauld's axle will turn one millstone [with a square hole, installed at the square end of the axle] but not the other [with a round hole, installed at the round end of the axle]' (1991: 168).

In view of the extensive metaphysical and epistemological ramifications of Ayers's point, my conclusion will have to be tentative. I suggest that the most hopeful response to the justification challenge would be one that blends naïve realism with the rationalist insight that mechanical interactions have a distinctive intelligibility.[10] Such a response would enable us to avoid both Strawson's dispositional view of shape concepts and the view he is concerned to resist, that knowledge of the dispositions associated with shape properties amounts to possession of an empirical theory.[11]

References

Anscombe, Elizabeth (1993) 'Causality and Determination'. In: *Causation*, E. Sosa & M. Tooley (eds), pp. 88–104. Oxford: Oxford University Press.

Ayers, Michael (1991) *Locke*, vol. ii. London: Routledge.

Campbell John (2002) *Reference and Consciousness*. Oxford: Oxford University Press.

Hume, David (1975) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals, reprinted from the 1777 Edition*, ed. by L. A. Selby-Bigge; third edition revised by P. H. Nidditch. Oxford: Clarendon Press.

[9] Quoted in Ayers, 1991: 135.

[10] Campbell's account of the relation between perceptual experience of shape properties and our grasp of the dispositions they ground can be read as one way to develop the rationalist insight. (See Campbell, 2002: ch 12.) On Campbell's account, the experience of a categorical property causes and justifies a subject's dispositions to engage in certain kinds of imagistic reasoning, where the latter are taken to be partly constitutive of the subject's knowledge of the dispositions objects have in virtue of the categorical property.

[11] I would like to thank Christoph Hoerl and Jim Woodward for very helpful comments on earlier versions of this chapter.

Leslie, Alan & Keeble, Stephanie (1987) 'Do six-month-old infants perceive causality?', *Cognition* 25: 265–88.

Putnam, Hilary (1975) 'Philosophy and our Mental Life'. In his *Mind, Language, and Reality* (*Philosophical Papers* vol. ii). Cambridge: Cambridge University Press.

Roessler, Johannes (2011) 'Causation in Commonsense Realism'. In: *Perception, Causation, and Objectivity*, J. Roessler, N. Eilan, & H. Lerman (eds), pp. 103–120. Oxford: Oxford University Press.

Schlottmann, Anne (2001) 'Perception Versus Knowledge of Cause and Effect in Children: When Seeing is Believing', *Current Directions in Psychological Science* 10: 111–15.

—— & Shanks, David (1992) 'Evidence for a Distinction between Perceived and Judged Causality', *The Quarterly Journal of Experimental Psychology* 44A: 321–42.

Shoemaker, Sydney (2003) 'Causality and Properties'. In his *Identity, Cause, and Mind*, pp. 206–33. Oxford: Oxford University Press.

Snowdon, Paul (2011) 'Perceptual Concepts as Non-Causal Concepts'. In: *Perception, Causation, and Objectivity*, J. Roessler, N. Eilan, & H. Lerman (eds), pp. 121–38. Oxford: Oxford University Press.

Strawson, Peter. F. (1988) 'Perception and its Objects'. In: *Perceptual Knowledge*, J. Dancy (ed.), pp. 92–112. Oxford: Oxford University Press. (First published in *Perception and Identity: Essays presented to A.J. Ayer*, G. McDonald (ed.). London: Macmillan, 1979).

—— (1992) *Analysis and Metaphysics*. Oxford: Oxford University Press.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

—— (2007) 'Interventionist Theories of Causation in Psychological Perspective'. In: *Causal Learning: Psychology, Philosophy, and Computation*, A. Gopnik & L. Schulz (eds.). Oxford: Oxford University Press.

—— (2011) 'Causal Perception and Causal Cognition'. In: *Perception, Causation, and Objectivity*, J. Roessler, N. Eilan & H. Lerman (eds.), pp. 229–63. Oxford: Oxford University Press.

4

Counterfactual and Other Forms of Conditional Reasoning

Children Lost in the Nearest Possible World

Josef Perner and Eva Rafetseder

"I'm lost, did you happen to see a lady without a boy like me?"

Source: www.CartoonStock.com

1 Introduction

Our naïve, driving question is: when do children become capable of counterfactual reasoning? Existing data do not provide a clear answer to our question. To shed some new light on this issue we analyse different kinds of conditionals and ask what kind of reasoning is required to understand them. Although we develop a systematic classification of several different kinds of conditionals and types of reasoning, in the end we focus on *counterfactual reasoning* required by subjunctives about the past and *basic conditional reasoning* for (timeless) conditionals typically expressed in the present to capture recurring regularities. In both cases assumptions counter to fact can occur (I can simply think of some event that I know is not going on). The developmental literature has focused on children's difficulty in reasoning with assumptions or thoughts that run counter to known facts which persists to about 4 or 5 years of age. But this ability does not differentiate between basic conditional reasoning with imagined (counterfactual) events or states of affairs[1] and counterfactual reasoning. In our analysis, when background assumptions are needed, basic conditional reasoning draws on plausible assumptions while counterfactual reasoning has to take such assumptions from the actual sequence of events (nearest possible world) to which it is supposed to be a counterfactual. When controlling for this factor children seem oblivious to this requirement until the age of 10 or beyond. These age limits may tie in well (once necessary controls have been implemented) with children's ability to experience counterfactual emotions like regret and relief.

The existing developmental literature locates the onset of counterfactual reasoning at different ages ranging from 2½ to 6 years. For instance, Harris (1997) reports that even children younger than 3 have a notion of counterfactuality in their understanding of 'almost' (but see Beck & Guthrie, in press). Harris, German, & Mills (1996) report that most 3-year-olds can answer explicit counterfactual questions correctly, while data from Riggs et al. (1998) point to the age of 4. There is even a suggestion in the air that children do not show regret until about 6 years of age (Amsel et al. 2003; Amsel & Smalley, 2000) because regret is an emotion triggered by counterfactual considerations. In order to be able to shed some light on these discrepant claims we have to first get clear on conceptual and methodological issues. How does counterfactual reasoning differ from other kinds of reasoning, and what would one need to show in order to be able to conclude that children are capable of counterfactual reasoning?

Counterfactual reasoning comes down to reasoning with counterfactual conditionals. Most developmental research assesses children's ability to make a prediction on the basis of a stipulated antecedent, e.g.: If there had been no fire, where would Peter be? (Riggs et al. 1998), although there is also some research on children's ability

[1] In developmental investigations mostly events are used. For conciseness, therefore, we use *event* to stand in for *situations* and *states of affairs*.

to explain what one should have done in order to prevent an actual outcome (Guajardo & Turley-Ames, 2001; Guajardo & Turley-Ames, 2004; Harris et al. 1996). We focus our discussion on making predictions and start by looking at how counterfactual conditionals have been placed in relation to other conditionals in the philosophical literature. From this we develop a distinction between different kinds of conditional reasoning of which, for our developmental considerations at the end, only the contrast between counterfactual reasoning and basic conditional reasoning will be central.

2 Classification of Conditionals

Counterfactual conditionals are evidently a species of conditionals. Conditionals are linguistically expressed by 'if-then' statements. The complement of the if-clause is called the *antecedent*, and the complement of the then-clause the *consequent*. They are typically classified by the linguistic mood of these complements as *indicative* and *subjunctive conditionals*. The subjunctive conditionals are also called *counterfactual conditionals* (Edgington, 2006). Another relevant feature under discussion is the relative time to which the antecedent and consequent are intended to refer: to no particular time, to the past, present, or future. Table 4.1 illustrates the resulting distinctions by adopting a slightly changed example from Stalnaker (1975), adapted for counterfactuals by Edgington (1995). The table also makes clear that the intended temporal reference can be glimpsed from the tense, and that conditionals that are not intended to refer to any particular time can be expressed in any tense.

The most natural way to express conditionals without intended time reference is to use the indicative present or future. We thus will only use the indicative present (or future) to describe cases of type **A.**

Table 4.1 Classification of conditionals

Intended temporal reference	Tense	Linguistic mood	
		Indicative	Subjunctive
none	any (typically present)	**A** If (whenever) the gardener doesn't do it, no one does it.	**A′** If (on any occasion) the gardener weren't doing it, no one would.
		B	**C**
past	past	If the gardener didn't do it, no one did.	If the gardener hadn't done it, no one would have.
present	present	If the gardener doesn't do it, no one does.	If the gardener weren't doing it, no one would be.
future	future	If the gardener won't do it, no one will.	If the gardener weren't to do it, no one would.

Most of the discussion of conditionals with a specific time reference has focused on conditionals that refer to the past or present where the linguistic mood makes an important difference. Indicative conditionals (**case B**) can differ from subjunctive (counterfactual) conditionals (**case C**) in truth value as the classic example from Adams (1970) shows.

Someone who doesn't believe in any of the conspiracy theories about Kennedy's assassination may well agree to the statement, 'If Oswald did not kill Kennedy, someone else did' (because we accept for a fact that Kennedy was killed and someone must have done it). Yet the same person may quite rationally object to the same statement only expressed in the subjunctive: 'If Oswald hadn't killed Kennedy, someone else would have' (because, one would have to strongly believe in a conspiracy theory and assume that if Oswald had failed to kill Kennedy then someone else would have jumped in and done the job). The consensus seems to be that this difference in mood is the same for reference to past or present. There is controversy as to whether this difference between moods persists for reference to the future (Bennett, 2003; Edgington, 2006; Morton, 2004). For our purposes we use mostly examples for the past and have labelled them as **cases B** and **C** in our table.

Before we can continue we need to know what the 'its', the one in the antecedents and the frequently omitted one in the consequents, refer to. Let us assume they refer to *the picking of the roses for the lady of the manor* (or perhaps it was *the killing of the Reverend*—but let's stick to roses). Whether we can agree to the various conditionals depends on the unwritten laws of the manner or the ways things are usually run there. These regularities provide us with the conditionals for our reasoning. Let's say I am the new manager being introduced by you to the workings of the manor. You just explained to me that:

(1) It is strictly the gardener's duty to pick the roses for the lady before 8 a.m. If he doesn't pick them, then no one else does. If nobody picks roses, then there aren't any on the table. If there are no roses on the table by 9 a.m., the lady is ill tempered.

Now we go to the dining room at 8 a.m. and see the roses on the table and you suggest a little test of my rule knowledge (**case A**) with the question: 'What if the gardener doesn't pick the roses?' You are not referring to any particular incident in the future, past or present, but just any assumed case. What is required is *basic conditional reasoning* (Rafetseder, Cristi-Vargas, & Perner, 2010). In this case, it would be using modus ponens repeatedly:

(2) If the gardener doesn't pick the roses, then nobody picks the roses.
The gardener doesn't pick the roses.
Therefore: Nobody picks the roses.

With this conclusion as a premise for the next round, together with the rule that 'if no one picks the roses, then there aren't any on the table', I can answer, 'There won't be any roses on the table', and with a further inference of this kind: '. . . and the lady will be ill-tempered'.

A slightly different scenario: You point to the roses on the table and ask (**case B**), 'What if the gardener *didn't pick* these roses?' In this case I will be at a loss and must improvise, e.g. 'Then someone must have helped out'. In contrast, if you ask (**case C**): 'What if the gardener *hadn't picked* these roses?', then my answer would be, 'Then they wouldn't be on the table', etc. These three cases illustrate several important points.

(i) Very similar questions can trigger different kinds of reasoning. In **case A** we ignore the fact that the roses are on the table and the gardener must have picked them. In **case B** we question the presumed fact that the gardener must have picked the roses. In **case C** we acknowledge that the gardener did pick them but entertain an alternative course of events where the gardener didn't pick them.

(ii) Sometimes different kinds of reasoning lead to different answers, e.g. 'someone must have helped out' (**case B**) vs 'no roses on the table' (**case C**), and sometimes to the same conclusion (**case A** and **C**) that there are no roses on the table and the lady is ill-tempered. And this can change as the rules of the manor change. For instance, if the improved rules include:

(3) If the gardener doesn't pick the roses by 8 a.m., then the butler picks them before 9 a.m.

Again we see the roses on the table by 8 a.m. but we also received a sick call from the butler. Now you ask:

Context A (to check my knowledge): 'What if the gardener doesn't pick the roses?' and I answer, 'Then the butler picks them'. (Standard new rule—does not take specific circumstance of particular case, i.e. butler's sick call, into account.)

Context C (worrying about today's events): 'What if the gardener hadn't picked the roses?' Then I will answer, 'Then there wouldn't be any roses on the table'. (Refers specifically to today's instance—takes butler's sickness into account.)

Since I answered the subjunctive question differently from the indicative question we can tell from my answer to the subjunctive question that I must have used counterfactual reasoning and not basic conditional reasoning, as I did for the indicative question.

(iii) Unless one makes sure that different problems requiring different reasoning strategies lead to different answers, one cannot infer from the answer which kind of reasoning was used. In the original example, one cannot conclude from the subjunctive question being answered 'no roses on the table' that counterfactual reasoning was being used, since the same answer would also have been given by using basic conditional reasoning.

(iv) In all three cases I had to reason from, so to speak, 'counterfactual' premises to 'counterfactual' conclusions, i.e. from 'the gardener didn't pick the roses', when in fact he did, to 'nobody picked the roses', when in fact someone did, etc. This illustrates the importance of not equating involvement of 'counterfactual assumptions' (i.e. assumptions counter to fact) with 'counterfactual reasoning'.

3 Types of Reasoning

We have seen that depending on context, and with minor variation in tense and mood, otherwise very similar questions can trigger quite different reasoning processes. We now try to give a more systematic overview of these different reasoning processes. Our exposition will be subject to the following two restrictions:

(i) We only cover reasoning from antecedents to consequents (also called 'forward reasoning' in the developmental literature; see Guajardo & Turley-Ames, 2004) as it can be covered by modus ponens[2] (see (2) above). The relevant distinctions we draw will, we think, also apply to modus tollens ('backward reasoning') and other forms of logical inference.

(ii) We assume that the information given by the conditional reflects an actual regularity in the world that one reasons about.[3] Only the second premise, which instantiates the antecedent, we allow to be non-factual, i.e. merely imagined or assumed.

To explain the different types of reasoning, let us review some necessary terminology. The world we reason about is the *target world*. This is often the real world but need not be. I can reason about the world of Santa Claus or other fiction. The elements of our target world can form different states and events (or sequences of events) that extend over time. The laws (regularities) that reign in the world determine which states and events are possible. Although many events are possible at a particular time, only a few actually take place; these we call *actual events*. We get to know about actual events through information (perception, testimony) about them, and by reasoning using the laws of the target world. So, if I hear at 9 a.m. that the gardener has picked the roses, I can know by inference that the roses are on the breakfast table without having seen them.

The pieces of knowledge we gain in these ways about the actual events we can combine in a *mental model* of the actual events. We can also build mental models of other possible event sequences, *fictive (imagined) events* that are not actual events. We simply think of or assume some event taking place and then use the laws of the world to reason about what other events must or will take place.

[2] As mentioned earlier, because developmental research has focused on children's ability to make predictions, we restrict our focus on the reasoning from antecedents to consequents.

[3] At a higher level, of course, one can also hypothesize (as in much of science) and make counterfactual assumptions about regularities (conditionals), e.g. even though it is true that if the gardener does not pick the roses no one else will, one can still assume counterfactually what would have happened if different rules were operating at the manor and the butler would always step in for the gardener.

With these distinctions in place we can start specifying the differences between the different types of reasoning. They can be captured by two factors: (i) whether the antecedent-instantiating premise represents an actual or imagined event, and (ii) how a model of an imagined sequence of events relates to actual events. Reasoning where only the first factor comes into play we call *Basic Conditional Reasoning* as it only requires a single model to reason with. When the second factor becomes relevant we speak of *Complex Conditional Reasoning* which needs two models and relating them to each other.

Basic Conditional Reasoning

1. *Reasoning about Actual Events*. We have knowledge of an actual event in our model and add further actual events to the sequence by applying the laws of the world. For instance, we hear at 9:30 a.m. that the gardener is sick and can't pick the roses. Knowing the rule 'if the gardener doesn't pick the roses, then no one will' and other rules, we can add to our model the further facts that there are no roses on the table and that the lady is ill-tempered.
2. *Reasoning about Imagined Events*. We need no knowledge of what actually happened. We simply imagine or assume some fictive event happening and add further events to the sequence by applying the laws of the world, e.g. we imagine the gardener having been sick all day and figuring out from that assumption that there aren't any roses on the table and that the lady is furious.
3. *Reasoning with Imagined Extensions of Actual Events*. This type of reasoning mixes reasoning about actual and imagined events. An example frequently mentioned in the developmental literature is *future hypothetical reasoning* (Riggs et al. 1998). In terms of our lady's manor, it might run like this: We receive at 8 a.m. a call from the gardener that he can't come in. So we put the fact that the gardener has not picked the roses in our model. Continuing from this fact in our model we can now start imagining what will happen if we assume that we ask the butler to kindly step in for the gardener before 9 a.m., that he will pick the roses, and the lady will be cheerful.

It is a developmentally important point about these three types of basic conditional reasoning that the reasoning only needs a single model. In particular, even in the mixed cases of category 3 the actual events can be combined with the imaginative projection of events into the future within a single model. And indeed, these types of reasoning have been found to be mastered early on by children (Beck et al. 2006; Perner, Sprung, & Steinkogler, 2004; Riggs & Peterson, 2000; Riggs et al. 1998; Robinson & Beck, 2000, Study 1). The sufficiency of a single mental model in these cases contrasts with the types of reasoning described as complex conditional reasoning where more than one model needs to be considered.

Complex Conditional Reasoning

4. *Counterfactual Reasoning*. Here we need to build a model of a sequence of imagined events but not just any sequence. It needs to be one that is taken counter to an actual sequence of events (hence 'counterfactual'). The initial, imagined event is typically the negation of an actual event, e.g. 'if the gardener hadn't picked the roses' implies that he did do it because of use of the subjunctive.[4] In contrast to basic reasoning with imagined events, one cannot further elaborate the model with further fictive events purely by applying rules of the target world. The model to be developed from the imagined counterfactual event has to obey a further constraint: it has to be modelled as closely as possible on the actual sequence of events, e.g. to the sequence of events in which the gardener picked the roses as implicated by the counterfactual antecedent 'If the gardener hadn't picked the roses'.

   Logicians tried to capture this constraint with the notion of a 'nearest possible world' (Stalnaker, 1968) or a set of such worlds (Lewis, 1973).[5] A counterfactual conditional is true if the corresponding conditional is true in a possible world in which the antecedent and the consequent of the conditional are true and which is otherwise as similar as possible to the real world (there is no possible world more similar to the real world, in which the conditional is not true). This proposal has met problems in giving a coherent account of the term 'nearest' (Menzies, 2008).

   Fortunately, for our developmental concerns we just need to capture the intuition behind these proposals, which we did in somewhat different terms. The important lesson to observe is to use counterfactual test problems in which basic reasoning with imagined events does not yield the same answers as counterfactual reasoning.

5. *Critical Reasoning*. Here we entertain different hypotheses about what actually happened in order to find out what actually happened. This is the reasoning that good detectives use, and critical minds and scientists should use, but at the higher level of discovering laws, where it is known as hypothetico-deductive reasoning. In other words, it is reasoning aimed at finding out more about the actual events (Morton, 2004: 292—to repair one's belief system). In order to bring out the critical difference with counterfactual reasoning we need to say some more about mental models and how they relate to actual and imagined events.

   Mental models are representations that represent actual or imagined events. What a representation represents is its *representational target* (or *object*). The way the target is represented as being is the *representational content*. When the content

[4] Although the counterfactual antecedent is typically false, it is not a defining characteristic of counterfactuals, as plausible examples by Edgington (1995) have shown.

[5] We thank Matthias Unterhuber for pointing out that Robert Stalnaker insists on a single nearest possible world while David Lewis admits a set of such worlds.

diverges from how the target actually is then we have a case of misrepresentation, e.g. if in my beliefs I represent the gardener as being short when in fact he is tall, then I have a false belief about his height. One can purposefully exploit this discrepancy between target and content underlying misrepresentation by entertaining different hypotheses about how something is. This ability is asked for in critical reasoning. Two (or more) models (representations) are constructed of the same actual sequence of events (representational target) but differing in content (showing the sequence as being different in each model). The critical part comes in by checking for each model whether the hypothesis is compatible with the more reliably known facts. If it isn't compatible, we deduce that the actual event could not be the way it was assumed to be in this model. For instance, we heard that the gardener was sick but we see the roses on the table. So who picked them? Obviously something happened today that contained an event of picking the roses on the table. We do not know who did the picking but we can entertain different hypotheses (alternative models) about what actually happened and then look for further evidence to decide between them.

4 Development: Task Demands–Solution Strategies

In the previous sections we worked out the differences between different kinds of conditional reasoning, in particular, between basic conditional reasoning and counterfactual reasoning. The use of the subjunctive in connection with the past tense signals most clearly a counterfactual problem for which counterfactual reasoning should be used. Basic conditional reasoning is typically called for by the indicative and any tense.

When we investigate children's ability to reason counterfactually by giving them counterfactual tasks, a central question is how we can tell whether a child actually brings to bear counterfactual reasoning on the task. Obviously this is done by checking whether they can give reliably correct answers on counterfactual problems. However, one has to be careful to choose the test problems so that any other reasoning strategy would not yield the same 'correct' answers. This precaution has not been observed in the developmental literature.

Basic Conditional Reasoning for Counterfactual Problems: False Positives. Explicit counterfactual questioning has relied on using the subjunctive in past tense. Although these questions pose problems for young children, there is no guarantee in existing studies that when they can give the correct answer they have engaged in counterfactual reasoning. Harris et al. (1996) told 3- and 4-year-old children short enacted stories. For instance, children observed a doll making dirty footprints on the clean floor and were asked the test question: 'What if Carol had taken her shoes off—would the floor be dirty?' 75% of the 3-year-old children and 87% of the 4-year-old children gave correct answers. The few errors that 3-year-olds made were so called 'realist errors', i.e. they answered in terms of the observed, real state of the scene with the dirty floor, and

answered 'yes'. For giving the correct 'no' answer, children could get away with basic conditional reasoning without consideration for what actually happened. This can be seen by the fact that one could, out of the blue, be asked the question, 'What if Carol had taken her shoes off—would the floor be dirty?', and one could answer with some confidence 'no' rather than 'yes'. So the data may show that most 3-year-olds can give answers that are counter to fact, but there is no strict evidence that they arrived at this answer by counterfactual reasoning.

In the stories by Riggs et al. (1998; and Robinson & Beck, 2000), the specific real event cannot be ignored. For instance, children are told of Peter the firefighter who feels sick and goes to bed. Suddenly the fire sirens sound. Being an extremely conscientious fireman, Peter, despite feeling sick, reports at the sight of the fire. When asked 'If there had been no fire, where would Peter be?', only 41% of 3½- to 4½-year-old children answered correctly, noticeably fewer than in the study by Harris et al. (1996). One evident difference between the studies is that if one were asked without hearing the story, 'Where would Peter be if no fire had broken out?', one would have no clue what to say.

This difference between tasks may indeed be important. In a study by German & Nichols (2003) children were presented with stories involving four events that were causally linked, e.g. Mrs. Rosy is very happy because she has planted flowers in her garden. (1) She calls her husband to come and see. (2) When her husband opens the door, the dog escapes from the kitchen. (3) The dog runs around and jumps on the flowers. (4) Mrs. Rosy is very sad about the squashed flowers. Children found it easy to make a short chain inference: 'What if the dog hadn't squashed the flowers, would Mrs. Rosy be happy or sad?' (3-year-olds were 69% correct). However, 3-year-olds gave few correct answers for medium (~26%) and long chain (15%) inferences: 'What if the dog hadn't escaped from the house/Mrs. Rosy hadn't called her husband, would Mrs. Rosy be happy or sad?', while 4-year-olds performed quite well (100%, 86%, and 67% for each length of chain, respectively). Although the authors interpret this difference in terms of the length of the inference, it also makes sense from our point of view. The question for the short chain ('What if the dog hadn't squashed the flowers, would Mrs. Rosy be happy or sad?') can be answered with some confidence with 'happy' even without hearing the story, while the other questions could not.[6]

In fact, recent evidence by Beck, Riggs, & Gorniak (2010) suggests that length of inference chain does not account for children's difficulty. In two studies they did not find any difference in performance at all, and in two further studies they found even the

[6] And even the minor difference between medium and long chains makes sense. One could argue that 'a dog escaping' has some negative connotation. Hence, when the question concerns the assumption that *the dog hadn't escaped* it might seem more likely that Mrs. Rosy would then be 'happy' than when the question is about *her not calling her husband*. This could also explain why there were more 'happy' answers to the medium than the long causal chains.

opposite pattern that counterfactual questions about short chains were more difficult than about long chains.

Rafetseder and Perner (2010) therefore looked at the other factor that could account for German & Nichols' original finding, namely, that counterfactual questions that can be answered independently of the sequence of actual events in the story are easier than those that need to refer to the actual events. In one of their stories a doctor is either working in the hospital (stereotypical location for doctors) or sitting in the park reading a book (atypical location). He gets an emergency call because a boy had an accident at the swimming pool. The doctor hurries to the swimming pool and is now treating the boy there. Children are asked: 'What if the boy hadn't hurt himself, where would the doctor be?' The idea is that when asked the question without being told the story (or listening to the story) one might answer more likely 'hospital' than 'park'. And indeed, especially 3½-year-olds gave more 'correct' answers in the stereotypical (86%) than the atypical condition (64%).

Basic Conditional Reasoning + Default Assumptions. We have argued that counterfactual reasoning differs from basic conditional reasoning with fictive events in terms of how it relates to the specific actual sequence of events, to which it is meant to be counterfactual. One cannot ask a subjunctive question about the past without mentioning a specific event. In both cases, for basic conditional and counterfactual reasoning one has to construct a model in which the antecedent is true and then look at what else is true within that model. This needs background assumptions. In basic conditional reasoning, the needed assumptions are made by default, i.e. of whatever is plausible or any suitable assumption that happens to come to mind. In counterfactual reasoning all the details of the actual sequence of events, to which the antecedent is taken to be counterfactual, have to be imported into that model. Now, we saw that in those tasks (e.g. Harris et al. 1996; short chain inference in German & Nichols, 2003; stereotypic alternative in Rafetseder & Perner, 2010), which are mastered by most 3-year-olds, basic conditional reasoning with fictive events (ignoring the actual sequence of events) leads to the correct answer. But in the tasks that pose more difficulty for these children (Riggs et al. 1998; medium and long inferences in German & Nichols, 2003; atypical location in Rafetseder & Perner, 2010), the questions asked could not be answered correctly without taking into account what actually happened ('If there had been no fire, where would Peter be?' can only be answered if one knows that he actually was and intended to stay at home). By the age of 4 most children also give correct answers in those tasks. So, it looks as if by about 4 years of age children are able to reason counterfactually.

Before agreeing to this conclusion we still need to examine the reason why children answer in terms of the actual event. For instance, do they simply answer the question about Peter with 'at home' because they realize that when there is no fire Peter is unlikely to be in a place to which he went only because of the fire. The one other place that comes to mind, as it was mentioned in the story, is Peter's home. So children, who engage in basic conditional reasoning, would still use this as a suitable answer simply because it

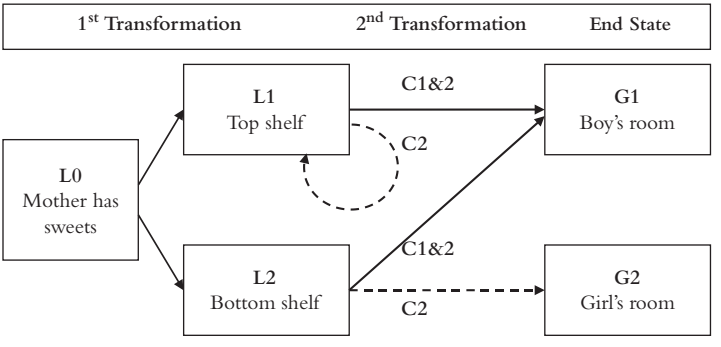


Figure 4.1 Structure of the story worlds (scenarios): sweets scenario example.
Source: Pilz (2005).

suggests itself as the only possibility and not because they obey the constraint of counterfactual reasoning, i.e. that the counterfactual series of events must be modelled as closely as possible on the actual series of events. So, we still don't know whether children's correct answers in these studies reflect genuine counterfactual reasoning or merely lucky hits on appropriate background assumptions for basic conditional reasoning.

Counterfactual Reasoning: Plausibility vs Actuality. To test whether children tend to give correct answers because the story material happens to suggest the correct background assumption for counterfactual reasoning or whether children see which assumption is asked for, Pilz (2005) devised stories with two-event sequences in which each event could be taken counterfactually. Stories were framed within different scenarios that all had the particular structure shown in Figure 4.1 for the sweets scenario. Whenever mother buys sweets she puts them either in a box on the *top shelf* (L1) or on the *bottom shelf* (L2). Then one of her two children comes looking for the sweets. It is either *her tall son* or *her little daughter*. The boy can reach both shelves (C1&2: character who can reach location L1 as well as L2), but the girl (C2) only the lower location (L2). So, if the boy comes he will take the sweets to his room, wherever mother has put them. Whereas the girl will bring them to her room only when mother has put them on the lower shelf (otherwise, they remain on the top shelf).

After familiarizing children with the regularities of this world and checking that they could reason about each eventuality, one of the four possible event sequences was actually played out and a counterfactual question was asked about the boy or the girl. For instance, the mother puts the sweets on the top shelf, the tall boy comes to collect them and carries them back to his room. 'Where would the sweets be if not the boy but the little girl had come to look for the sweets?'. This particular question proved to be very difficult: only two of seventeen 4;5-to-5;9-year-olds (*n*;*m* means *n* years and *m* months) gave the correct answer that the sweets would still be on the top shelf. The

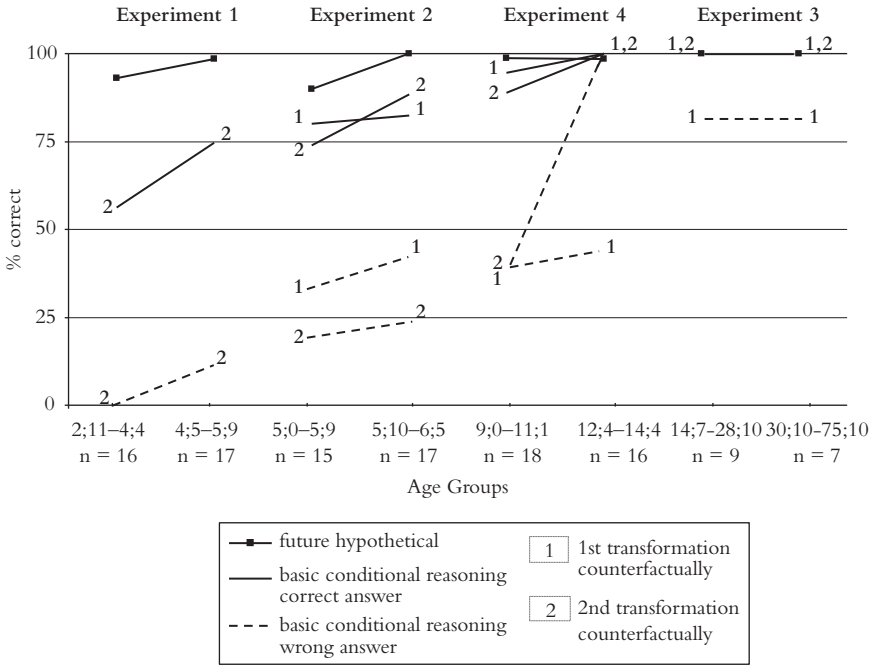


Figure 4.2 Percentage of correct answers to counterfactual questions depending on age.

Source: Experiments 1–3 taken from Rafetseder et al. (2010); Experiment 4 taken from Rafetseder & Perner (2009).

typical error was to say, 'in the girl's room'. In contrast, when asked directly what happens to the sweets on the top shelf when the girl comes to collect them, 94% correctly answered that they would stay on the top shelf. Moreover, 80% gave correct answers to the counterfactual question for the three other possible event combinations, where the answer was either that the sweets would be in the boy's room or in the girl's room. The data from this and three successor experiments are summarized in Figure 4.2.

In a slightly changed version of this experiment (Rafetseder et al. 2010, Experiment 2: the two characters' ability to reach locations was made symmetrical) slightly older children of 5;0-to-6;5 confirmed this pattern. Only 42% of these children answered the counterfactual question about the girl correctly with 'sweets would be on top shelf', but 92% did so when the correct answer was 'in her room'.

These findings raise the suspicion that even most 6-year-old children use basic conditional reasoning to answer counterfactual questions. They treat the question 'Where would the sweets be if the girl had come?' like the question 'What happens to the sweets when the girl comes?' Strictly speaking, one can't answer this question without knowing where the sweets are presumed to be—top or bottom shelf. However, children uncritically just latch onto the interesting possibility, namely that the girl

will take the sweets to her room, and tacitly make the rest of their imagined event sequence fit that view.

Since such 'old' children still show little sign of proper counterfactual reasoning, we wondered whether adults might live up to our expectations of how these questions ought to be answered. Reassuring for our intuitions, when given the same tasks not one of sixteen adults gave the response that children gave on these counterfactual questions (Rafetseder, et al. 2010, Experiment 3). Rafetseder and Perner (2009: Experiment 4 in Figure 4.2) showed that adult levels of performance on this task are reached by the age of 13 (12;4 to 14;4: 100% correct), while at the age of 10 they still were not much above the level of 6-year-olds (9;0 to 11;1: 39% correct).

In their experiments, Rafetseder et al. (2010) also asked counterfactual questions about the first transformation in the target world of Figure 4.1, i.e. where mother put the sweets. In this case a few adults gave the same answers as most children did. For children, the difficult question was, for instance, 'Where would the sweets be if mother had put the sweets on the top shelf?', after mother had actually put the sweets on the bottom shelf and the girl had come and taken them to her room. Only 41% of the 6-year-olds answered correctly with 'top shelf', another 12% said incorrectly that the little girl would be able to reach the top shelf, and the remaining 47% said that the sweets would be in the boy's room. This answer also fits the theory that these children used basic conditional reasoning and focused on the desire of the characters in the story to get the sweets into their respective room. If mother puts the sweets on the top shelf, where the girl can't reach them, then it would be more sensible to have the boy come and get them.

It is, however, unlikely that the few adults who gave the answer 'in the boy's room' did so for the same reason. They explained that they made the (reasonable) assumption that since the girl couldn't reach the sweets on the top shelf, they would stay there (the intended correct answer) but sooner or later the boy would come by, find them there, and bring them to his room. Perhaps some of the children also reasoned this way. This could explain why many (44%) of even 13-year-olds in Rafetseder & Perner's study (2009: see Experiment 4 in Figure 4.2) gave this answer, when the same children answered the difficult counterfactual question about the second transformation ('if not the boy but the small girl had come to get the sweets on the top shelf?') with 100% accuracy.

Finally, there is an interesting discrepancy between children's ability to engage in counterfactual reasoning in Rafetseder et al.'s (2010) sweets scenario and in the doctor story used by Rafetseder & Perner (2010), where by the age of 6, 60% of children consistently give correct answers to counterfactual questions. In the most difficult condition, the doctor sat in the park to read his book when he got an emergency call to the swimming pool, whereupon he went to the hospital to get his first aid kit and then to the swimming pool. This was thought difficult for children, who use basic conditional reasoning, because when asked where the doctor would be if there had been no accident, they simply pick the location that comes to mind first, i.e. the hospital, because doctors tend to be in hospitals rather than in parks and because

that's where the doctor was last before going to the swimming pool. Indeed, the hospital had a demonstrable attraction for children even at the age of 6. However, many children at this age did opt for the park as their answer. Did they understand the constraint of counterfactual reasoning to keep the counterfactual sequence of events as close as possible to the actual event, or could there be another reason? The, retrospectively obvious, possibility is that these children (not the very young ones below 4 or 5 years of age) registered when listening to the story that the doctor wished to sit in the park and read his book. When asked where the doctor would be if no accident had happened, they remembered his existing desire (an actual state of the mind) to be in the park and answered with this location. No special counterfactual reasoning constraint needed to be operating.

In sum, from these data and considerations it looks like children do not reason counterfactually by obeying the counterfactual reasoning constraint of staying as close as possible to the actual sequence of events until they are at least 6 years of age or older. The younger ones sometimes appear to follow this constraint but plausibly do so only for extraneous reasons, i.e. that the telling of the actual events makes these events more likely to be used as defaults in constructing the counterfactual model. This conclusion seems surprising in view of earlier claims but is compatible, since these earlier claims were based on data that did not control for basic conditional reasoning strategies. It also creates an interesting developmental basis for children's susceptibility to counterfactual emotions such as feeling regret and relief.

5 Regret and Relief

Amsel et al. (2003) assessed pre-schoolers', 2nd graders', 5th graders', and college students' emotional reactions in a card game. Participants had to turn over one of two cards lying upside down. Trials of interest were those where the chosen card tied with the experimenter's card lying the right side up. Participants had to rate (+1 happy, 0 neither, −1 sad) their satisfaction with their choice (*Initial Rating of Chosen Card*). Then the unselected card was revealed, which in one case would have trumped the experimenter's (positive alternative outcome) and in the other case would have lost against the experimenter's card (negative alternative outcome). Participants were asked to rate their feelings if they had chosen the alternative (*Rating of Alternative*). With the exception of one 5th grader, all participants judged themselves as happier if they had turned over a winning card and less happy with the losing card than they were with their tie. The mean absolute difference between *Initial Rating* and *Rating of Alternative* was between .50 and .57 for all three age groups of children and .90 for adults. That is, the pre-schoolers (3½–5½-years-old) showed the same difference in rating between their actual choice and the alternative choice as the older groups.

The game then continued. The alternative card was turned upside down again and participants were asked once more about their chosen card: 'How do you feel *now*

about the card you turned over?' All adults now rated the satisfaction with their tying card as higher (relief) if the alternative had been a loser, and lower (regret) if the alternative had been a winner. And so did most of the older children: eleven of sixteen 5th-graders, thirteen of fifteen 2nd-graders, but only five of fourteen pre-schoolers (and one showed the opposite tendency). This is also reflected in the average rating differences: .17, .56, .62, and .90 from pre-school to adulthood, respectively. Similar results were reported by Beck & Crilly (2009).

The interesting fact here is that the pre-schoolers showed hardly any regret/relief effect even though their ratings of the alternative card were very much the same as that of their elder peers. This is theoretically interesting because the rating of the alternative outcome can be based on basic conditional reasoning within a single model about a fictive counterfactual sequence of events: 'if I get a winning/losing card then I feel happy/sad'. In contrast, they rated their feeling about the tie as 'neutral'. In contrast, the change in the rating of the chosen card, upon discovering what the alternative would have been, seems impossible within a single model. The model showing the counterfactual sequence of fictive events must influence the rating given on the basis of the model for the actual events.

Unfortunately, this interesting developmental trend is based on the problematic methodology of repeated questioning. For instance, in the regret condition, when asked first to rate their satisfaction with the tie, most will say 'neutral'. If they are then shown the winning alternative most will say that they would be 'happy' with that card. Then they are asked about their actual card again, to which they had responded with 'neutral'. Being asked the same question a second time has the tacit implication that a different response is in order. They can't switch to 'happy' since this has been used for the clearly better winning alternative, so they switched to 'sad' instead, and it appears that they experienced regret, when in fact they were just trying to avoid giving the same response twice.

To avoid this problem of repeated questioning Rafetseder & Perner (submitted) simplified the procedure. Children had the choice between one of two boxes to keep its content for a reward. The chosen box contained one candy. Children in the 'before' condition were asked how happy they were about receiving a candy. Children in the 'after' condition were first shown that the alternative box contained five candies before they were asked how happy they felt about their one candy. They gave their ratings by adjusting a fine grained, quasi continuous, barometer varying from a sad face (step 0) to a happy face (step 64). The results are shown in Figure 4.3. Pre-school and school children up to the ages of 8 to 9 showed practically no difference. Only by the age of 10 did some, and by the age of 13 did most children show a clear reduction in happiness.

Guttentag and Ferrell (2008) also found that children did not exhibit an understanding of anticipatory regret before the age of 10. In their study, participants ranged from 5 years of age to adulthood. They were shown three boxes, one of which contained a big prize, another a medium-sized prize, and the third had nothing in it. Participants were asked to select one box which was then removed by the experimenter. After that they

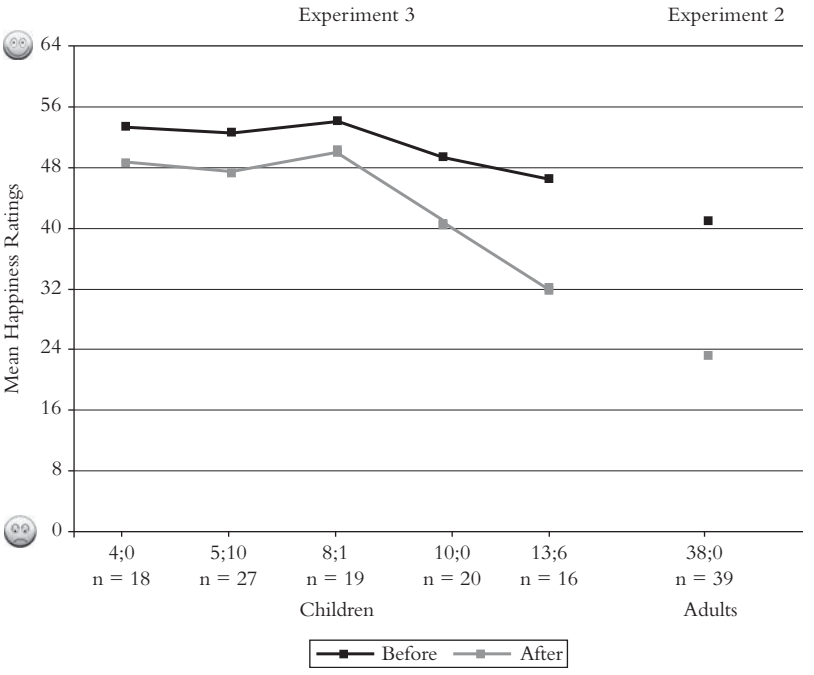


Figure 4.3 Mean happiness ratings before and after the alternative outcome was revealed.
Source: Data from Rafetseder & Perner (submitted).

had to choose one of the remaining two boxes. Participants were then shown that the selected box contained the medium-sized prize and had to rate how they felt on a five point facial-expression scale. Afterwards they were asked whether they wanted to see what was in the box they hadn't selected. Anticipation of regret would lead to the wish that there was nothing in that other box. Indeed, 44% of 9- and 10-year-olds and most adults (70%) had that wish. In contrast, only 5% of the 5- and 6-year-olds and 11% of the 7- and 8-year-olds expressed that wish. On the contrary, most of them said they wished the big prize were in that box.

This result nicely coheres with the findings from Rafetseder et al. (2010). Between the ages of 10 and 13, children become able to systematically tune their models of counterfactual events to their models of actual events and implement the 'nearest possible world' constraint of counterfactual reasoning. At about the same age, their emotional reactions to real events are modulated by their counterfactual considerations (Rafetseder & Perner, submitted). This neat picture is, however, compromised by the small but still systematic rating differences shown by the younger children. When all three age groups were put together the difference was statistically reliable. What we need to explain is why there is this small but steady difference over several years which then starts to increase by about the age of 9 and reaches adult levels by the age of 13.

A possible explanation is that our emotion ratings do not just depend on regret but also on the simple frustration of not having got 5 candies. In other words, the children in the 'before' condition rate their satisfaction in view of the actual event of having got a candy, while the children in the 'after' group rate their satisfaction of a mix consisting of having received a candy and of not having received 5 candies. If they do so, no wonder their ratings were lower in the 'after' than the 'before' condition. The overall picture, then, would be that the small rating difference shown by even the youngest is due to this frustration effect, and the increase of the rating difference starting around the age of 9 is due to feelings of regret being added.

An easy test of this explanation is to also use relief trials in addition to regret trials. If children receive one candy and learn that they would have got nothing in the other box then this should lead to counterfactual feelings of relief and a higher satisfaction rating in the 'after' than in the 'before' group. The inverse frustration effect is implausible, i.e. that children feel better for getting a candy and not getting no candy than for getting a candy. So the prediction is that on relief trials there should be no rating differences between the 'before' and 'after' conditions until children reach the age of 9 or 10, when their rating differences start to increase.

6 Conclusion

We analysed how counterfactual reasoning differs from other kinds of conditional reasoning, in particular, basic conditional reasoning about fictive events. These fictive/imagined events can contain propositions that run counter to fact (counterfactual elements). Counterfactual reasoning, therefore, is not just reasoning with such counterfactual elements. Rather, it needs to meet the 'nearest possible world' constraint of modelling the fictive events as closely as possible on the real sequence of events to which it is a counterfactual. The literature on children's ability to reason counterfactually has mostly focused on their problems in reasoning with antecedents that are counterfactual—a problem that dominates children's reasoning around the ages of 3 and 4. The few studies that have attempted to control for the nearest-possible-world requirement indicate that children do not obey this constraint until the age of 10 or later, which may tie in with evidence on their ability to experience counterfactual emotions (regret, relief) once necessary methodological precautions have been taken.

Younger children's errors on certain counterfactual problems suggest that they tackle such problems with basic conditional reasoning. But often they appear to reason counterfactually because the task conditions happen to present them with the correct actual events for constructing their counterfactual model. In other words, children often find themselves in the 'nearest possible world', but without meaning to be there (for proper counterfactual reasoning, however, one should be there on purpose).

References

Adams. E.W. (1970) 'Subjunctive and Indicative Conditionals', *Foundations of Language* 6: 89–94.

Amsel, E., Robbins, M., Fumarkin, T., Janit, A., Foulkes, S., & Smalley, J.D. (2003) 'The card not chosen: The development of judgments of regret in self and others'. Unpublished manuscript.

—— & Smalley, J.D. (2000) 'Beyond really and truly: Children's counterfactual thinking about pretend and possible worlds'. In: *Children's Reasoning and the Mind*, P. Mitchell & K.J. Riggs (eds), pp. 121–47. Hove, East Sussex: Psychology Press.

Beck, S.R. & Crilly, M. (2009) 'Is understanding regret dependent on developments in counterfactual thinking?', *British Journal of Developmental Psychology* 27: 505–10.

—— & Guthrie, C. (in press) 'Almost thinking counterfactually. Children's understanding of close counterfactuals', *Child Development*.

—— Riggs, K.J., & Gorniak, S.L. (2010) 'The effect of causal chain length on counterfactual conditional reasoning', *British Journal of Developmental Psychology* 28: 505–21.

—— Robinson, E.J., Carroll, D.J., & Apperly, I.A. (2006) 'Children's thinking about counterfactuals and future hypotheticals as possibilities', *Child Development* 77(2): 413–26.

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

Edgington, D. (1995) 'On Conditionals', *Mind,* 104(414): 235–329.

—— (2006) 'Conditionals'. In: *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). http://plato.stanford.edu/entries/conditionals/

German, T.P. & Nichols, S. (2003) 'Children's counterfactual inferences about long and short causal chains', *Developmental Science* 6(5): 514–23.

Guajardo, N.R. & Turley-Ames, K.J. (2001) *Theory of Mind and Counterfactual Thinking: Mutating the Antecedent versus the Consequent*. Poster presented at the biennial meeting of the Society for Research in Child Development, Minneapolis.

—— —— (2004) 'Preschoolers' generation of different types of counterfactual statements and theory of mind understanding', *Cognitive Development,* 19: 53–80.

Guttentag, R. & Ferrell, J. (2008) 'Children's understanding of anticipatory regret and disappointment', *Cognition and Emotion* 22(5): 815–32.

Harris, P.L. (1997) 'On realizing what might have happened instead', *Polish Quarterly of Developmental Psychology* 3: 161–76.

—— German, T. & Mills, P. (1996) 'Children's use of counterfactual thinking in causal reasoning', *Cognition* 61: 233–59.

Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.

Menzies, P. (2008) 'Counterfactual Theories of Causation'. In: *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.). http://plato.stanford.edu/entries/%20causation-counterfactual/

Morton, A. (2004) 'Indicative versus subjunctive in future conditionals', *Analysis* 64(4): 289–93.

Perner, J. (2000) 'About + Belief + Counterfactual'. In: *Children's Reasoning and the Mind*, P. Mitchell & K.J. Riggs (eds), pp. 367–401. Hove, East Sussex: Psychology Press.

—— Sprung, M., & Steinkogler, B. (2004) 'Counterfactual conditionals and false belief: a developmental dissociation', *Cognitive Development* 19: 179–201.

Pilz, R. (2005) *Kontrafaktualität und Wirklichkeit. Die Entwicklung zu kontrafaktischem Denken unter Einbeziehung der subjektiv bestehenden Wirklichkeit*. Unpublished Master's Thesis, Department of Psychology, Paris-Lodron-Universität Salzburg.

Rafetseder, E., Cristi-Vargas, R., & Perner, J. (2010) 'Counterfactual Reasoning: Developing a sense of "nearest possible world"', *Child Development* 81(1): 362–75.

—— & Perner, J. (2009) 'Counterfactual Reasoning: From childhood to adulthood'. Unpublished manuscript.

—— —— (2010) 'Is Reasoning from Counterfactual Antecedents Evidence for Counterfactual Reasoning?', *Thinking & Reasoning* 16(2): 131–55.

—— —— (submitted) 'When the Alternative had been better: Counterfactual Reasoning and the Emergence of Regret'. Under consideration by *Cognition and Emotion*.

Riggs, K.J. & Peterson, D.M. (2000) 'Counterfactual thinking in pre-school children: Mental state and causal inferences'. In: *Children's Reasoning and the Mind*, P. Mitchell & K.J. Riggs (eds), pp. 87–99. Hove, East Sussex: Psychology Press.

—— —— Robinson, E.J., & Mitchell, P. (1998) 'Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality?', *Cognitive Development* 13: 73–90.

Robinson, E.J. & Beck, S. (2000) 'What is difficult about counterfactual reasoning?' In: *Children's Reasoning and the Mind*, P. Mitchell & K.J. Riggs (eds), pp. 101–19. Hove, East Sussex: Psychology Press.

Stalnaker, R. (1968) 'A Theory of Conditionals'. In: *Studies in Logical Theory, American Philosophical Quarterly*, Monograph Series, 2. Oxford: Blackwell, pp. 98–112. Reprinted in F. Jackson (ed.) *Conditionals*. Oxford: Oxford University Press, 1991. Page references to the 1991 edition.

—— (1975) 'Indicative conditionals', *Philosophia* 5: 269–86.

5

Multiple Developments in Counterfactual Thinking

Sarah R. Beck, Kevin J. Riggs and Patrick Burns[1]

Mapping the development of children's counterfactual thinking should allow insight into this process in adults and potentially its relation with causal understanding. In this paper we argue that there is not one critical development that should be thought of as marking children's ability to engage in counterfactual thought, but rather a sequence of (at least) four developments taking place from early to middle childhood. We consider whether domain-general developments in executive function might underpin these developments.

To determine if a relation exists between counterfactual and causal understanding we need a description of the psychology of counterfactual thinking based on empirical evidence. We suggest that a series of important and impressive developments occur during childhood that together result in adult-like counterfactual thinking. We choose to take a developmental approach in order to obtain a picture of children's abilities at different ages and provide insight into the cognitive processes involved in adult counterfactual thinking.

We know that children start to answer explicit counterfactual conditional questions correctly at around 3 or 4 years of age. In many experimental tasks (e.g. Harris, German, & Mills, 1996; Riggs, Peterson, Robinson & Mitchell, 1998) children hear a short narrative and are asked a question containing an explicit false antecedent. For example, having seen the wind blow a picture into a tree children are asked 'What if the wind had not blown, where would the picture be?' When children make errors on these tasks they tend to be realist errors—that is, they describe the world as it is (the picture in the tree), not as it would be (the picture on the table). At about the same time children start to make references to counterfactual worlds in their spontaneous speech (Kuczaj & Daly, 1979) describing 'what if' something had happened. There is some evidence that

[1] This research was supported by grants from the Economic and Social Research Council, U.K. (RES 000 22 1683 and RES 062 23 1114 in collaboration with the E.S.F. Eurocores programme LogICCC).

children make implicit reference to counterfactuals much younger than this at 2-and-a-half when they start to use and understand the world 'almost' (Harris, 1997, but see Beck & Guthrie, in press) and when they engage in pretend play. Other evidence suggests that children's thinking is not adult-like until around 5 years of age (Beck, Robinson, Carroll, & Apperly, 2006). Due to these wide ranging findings, there is no clear consensus as to when children first start to think counterfactually.

In this chapter we focus on developments in children's explicit counterfactual thinking. Some authors suggest that causal understanding could be underpinned by implicit counterfactual thinking (e.g. Woodward, this volume). One claim that 14-month-old infants were employing implicit counterfactual thinking was made by Perner, Sprung, & Steinkogler (2004). They referred to Gergely, Bekkering, & Király's (2002) study in which infants learnt to use their head, not their hands, to work a light box if the model's hands were otherwise occupied. However, we are not convinced that this really demonstrates counterfactual thought, because it is not clear to us what evidence for implicit counterfactual thinking would look like. While one might describe the infant's behaviour in counterfactual terms, it is not clear whether the infant herself is in any sense thinking about a counterfactual. Thus, if one wants to explore the possibility of a relationship between causal and counterfactual thinking, it makes sense to turn to the evidence for explicit counterfactual thinking.

Our purpose in this paper is not to argue about the precise age at which children can answer counterfactual conditional questions. Instead, we suggest that the current focus in the developmental counterfactual literature has been too narrow. We present a review of the literature describing at least four types of thinking that may be considered to a greater or lesser extent 'counterfactual'. We outline evidence for a series of important developments that move the child towards a truly adult-like ability. These developments are: generating alternative worlds, representing falsity as if it were true, representing multiple possibilities, and comparing multiple possibilities. It may be that one of these developments is critical in children's causal thinking. Alternatively, it might be that each shift in children's counterfactual thinking has a parallel development in their causal thinking. We do not try to identify these relationships explicitly in this chapter, but by setting out these developments in children's counterfactual thinking we provide a framework that could serve this purpose.

We also consider the possibility that developments in domain-general cognitive processes contribute to this developmental progression. Unlike Sobel (this volume), who argues that children's counterfactual thinking is dependent on domain-specific knowledge about particular causal content, our main focus will be on the increasing evidence that domain-general executive functions underpin children's counterfactual thinking. Executive functions are those processes that govern goal-directed thought and behaviour and include inhibitory control, working memory, and attention shifting (among others). Inhibitory control can be viewed as the ability to ignore potentially interfering cognitions or actions that are irrelevant to the current task goal, working memory as the ability to maintain information in mind relevant to that task goal, and

attention shifting as the 'ability to change between mental sets or tasks' (Lehto, Juujarvi, Kooistra, & Pulkkinen, 2003). There are several reasons why we think developments in executive function and developments in counterfactual thinking are related:

1) We know there are substantial developments in executive function during early and middle childhood, particularly in those we think may be relevant for explaining counterfactual developments: inhibition, working memory, and attention shifting (see e.g. Brocki & Bohlin, 2004, Andrews-Espy, 2004).
2) A number of researchers have sought to establish the relationship between the development of executive functions and conditional reasoning in older children (Johnson-Laird & Byrne, 1991; Simoneau & Markovits, 2003; Handley, Capon, Beveridge, Dennis, & Evans, 2004). For example, Handley et al. (2004) found that inhibitory control and working memory predicted aspects of 10-year-olds' performance on logical reasoning task. The contributions depended on the content of the premises (i.e. whether it was true or false).
3) Theoretical accounts of counterfactual thinking imply a role for executive functions. Byrne (2005) proposed two aspects of counterfactual thought: holding multiple possibilities in mind and holding in mind 'a false possibility temporarily supposed to be true' (Byrne, 2007: 475). These abilities may be characterized as maintaining and evaluating information in working memory and resisting irrelevant information from known reality, which is likely to make inhibitory demands.

The four developments in counterfactual thinking we have suggested: generating alternative worlds, representing falsity as if it were true, holding multiple representations in mind, and making comparisons between possible worlds lend themselves to being described in terms of the executive functions. That is, they are easily translated into developments in inhibition, working memory, and attention switching. We will make specific predictions about which executive developments may be relevant for each type of counterfactual thinking, reviewing the empirical evidence where it is available. Note that if one does want to suggest a theory of causal understanding in terms of counterfactual thinking, then these domain-general developments are likely to be relevant to the development of causal understanding as well. It follows that any such account of the development of causal understanding should be of the generalist variety.

1 Creating Alternative Worlds

Besides pretend play where children clearly entertain fictional, non-real worlds, one of the earliest developments that shows that children can reason about an alternative real world is the ability to think about future hypotheticals. A test of future hypothetical thinking was devised by Riggs et al. (1998) as a comparison for their counterfactual conditional test. In one version of the experiment, children played a sorting game: pieces of paper with pictures on them were sorted into one tray, blank papers went into another tray. 3- and 4-year-olds were asked counterfactual questions after a picture had

been drawn on a piece of paper and it had been sorted, 'If I had not drawn on the piece of paper, which box would it be in?' and they were asked future hypothetical questions before a picture had been drawn, 'If I draw on this piece of paper, which box will it go into?' In both cases the child has to ignore the current state of affairs (whether or not the paper has a picture on it) and imagine an alternative. Yet there was asymmetry in children's performance. 3- and 4-year-olds found it much easier to answer the future hypothetical question than the counterfactual.

Robinson and Beck (2000) further investigated this. In their procedure a toy car drove along a road that had a garage at each end. When the car was in one garage children were asked, 'What if next time he drives the other way, where will he be?' (future hypothetical) or, 'What if he had driven the other way, where would he be?' (counterfactual conditional). Note that in this task the item is in one of the two possible locations (a garage) in both the future and counterfactual versions of the task. This was not the case in the original Riggs et al. experiment. In the future hypothetical version of their task the paper was on the table, not in either of the trays and this might have reduced the salience of the incorrect realist answer, improving performance. Despite this Robinson and Beck found the same pattern of results: 3-year-olds found questions about an alternative future event easier to reason about than an alternative to an event that has happened. The relative ease with which children answer future hypotheticals compared to certain counterfactuals has since been reported by Perner et al. (2004) and supporting evidence is reported by Beck et al. (2006).

Although we do not yet have a baseline for when the ability to consider future hypothetical worlds develops, the evidence suggests that from a very early age children can make sensible predictions about future events, ignoring the current conflicting state of affairs. Thus, it cannot be that simply ignoring current reality is what makes counterfactual conditional questions difficult for young children.

Before we leave this section we consider one interesting possibility. Perhaps this generation of alternative worlds as evidenced by future hypothetical thinking is underwritten by an understanding of causal concepts that Roessler considers in his chapter (Roessler, this volume). According to the particularist account that Roessler discusses, understanding particular causal transactions such as 'push' or 'knock over' may not require the ability to reason about counterfactuals. Rather, such simple causal concepts are a function of our perception of physical objects. This is contrasted with the acquisition of a 'general concept of causation', which, under an account such as Woodward's, only becomes available in tandem with counterfactual reasoning abilities. One possibility then is that the development of future hypothetical thinking is linked to an understanding of particular causal concepts which characterizes children's very early 'simple' causal reasoning.

Thinking about future hypothetical worlds is a first step in children's developing counterfactual thinking in that it involves thinking about an alternative world to the current one. However, it does not require one to negate what one knows to be true, a

key feature of counterfactual thinking according to social and cognitive psychologists. It is this kind of thinking that we consider next.

2 Truth and Falsity

What happens at about 4 years when children start to answer counterfactual conditional questions correctly? The difference between thinking about a counterfactual event and a future hypothetical event is that in the former case what is being imagined is known to be false. In the counterfactual conditional the child has to think about something that contradicts what she knows to be true. In the case of future hypotheticals she does not know the true state of affairs. There are several examples of tasks where children have to contradict what they know has happened in order to give the correct answer to a counterfactual question. For example, in Riggs et al.'s (1998) paper children heard unexpected transfer-type narratives, where an object or person moved from one place to another. In one of their stories Peter is at home in bed when he receives a phone call asking him to go help put out a fire at the Post Office. Once he is there the counterfactual conditional question is, 'If there had been no fire, where would Peter be?' Riggs et al. found substantial improvements between 3 and 4 years on these types of tasks. Children stopped giving realist answers, that Peter would be in the Post Office, and instead began to give counterfactual answers, that he would still be in bed.

Children's early success on future hypothetical questions suggests that their problem is not with generating new alternatives. In support of this Guajardo & Turley-Ames (2004) found that similar developments between the ages of 3 and 5 occurred in children's abilities to answer consequent counterfactual conditional questions, such as the one above about Peter and the fire, and in generating multiple counterfactual antecedents. For example, if a character walks through mud, comes in to the house, and makes the floor dirty it is possible to imagine many different answers to the question 'What could you have done so the kitchen floor would not have gotten dirty?' such as not walking through the mud, taking the boots off, wiping her feet.

There is an ongoing dispute about precisely when children can answer counterfactual conditional questions correctly. Harris, German, & Mills (1996) and German & Nichols (2003) report that 3-year-olds tend to give correct answers, whereas Riggs et al. (1998) and Guajardo & Turley-Ames (2004) claim that success is not seen until 4 years. Perner (2000; Perner et al. 2004) has suggested that children only find it difficult to answer counterfactual questions which involve thinking about the specifics of the events described (see also Perner & Rafetseder, this volume). He suggests that Harris' test stories could be answered based on background knowledge rather than by imagining a counterfactual alternative to the story. For example, one of Harris' stories was about a character, Carole, making muddy footprints on the floor as in the example in the previous paragraph. 3-year-olds found it relatively easy to answer the counterfactual question, 'What if Carole had taken her boots off, would the

floor be dirty?' Perner argues that one might answer this question by referring to general knowledge about whether floors get dirty when you do not have shoes on, which would not involve counterfactual thinking. A related question is whether some counterfactual questions might be easier for children to answer because of the content they involve. Sobel (this volume) argues that judgements about possible events, both counterfactual and future, are dependent on children's domain-specific causal knowledge. Although Sobel presents some evidence that children treat events in different domains differently, this factor cannot explain the differences between tasks that cause such controversy in the literature. Counter-evidence comes from the observation that in some studies the same causal content is required for counterfactual questions that prove easy (e.g. German & Nichols, 2003 short causal chains; Beck et al. 2006, standard counterfactuals) and others that are difficult (e.g. German & Nichols, 2003, long causal chains; Beck et al. 2006, open counterfactuals). Furthermore, in a study in which counterfactual questions about different causal domains (physical change and emotional change) were directly compared, no consistent differences were observed between them (Beck, Riggs, & Gorniak, 2010).

Rather than focussing on the age at which children can correctly answer these counterfactual conditionals, here we are interested in the progression of developing cognitive abilities. Most studies find that children's performance on these types of questions improves between 3- and 5-years (for an exception see German & Nichol's short causal chain stories, 2003, but see Chan & Hahn, 2007, and Beck et al. 2010). So what further cognitive abilities do children need to think about counterfactual alternatives rather than future hypothetical alternatives? Several authors have suggested that it is the conflict between reality and the imagined alternative which causes children difficulty (e.g. Peterson & Riggs, 1999; Perner et al. 2004). Evidence from our own recent study allows us to address this question (Beck, Riggs, & Gorniak, 2009).

We ran a battery study testing ninety-three 3- and 4-year-olds on a range of counterfactual and executive function tasks. We included two types of unexpected transfer-type counterfactual conditionals: location change stories (based on false belief unexpected transfer stories) from Riggs et al. (1996) which were acted out with toys and the short and long counterfactual emotion stories from German & Nichols (2003) which were illustrated with pictures. In one story based on Riggs et al. (1996), Piglet was painting a picture in the garden, he left it on the table, then the wind blew the picture up into the tree. We asked children, 'What if the wind had not blown, where would the picture be?' The correct answer was to say 'on the table'. In one of the German and Nichols (2003) stories, Mrs Rosy planted a flower in the garden and called her husband to come and see it. He opened the kitchen door, letting the dog escape, who then squashed the flower. Children were asked short causal chain counterfactual questions, 'What if the dog had not squashed the flower, would Mrs Rosy be happy or sad?' and long causal chain counterfactual questions, which required them to change an earlier event in the story, 'What if Mrs Rosy had not called her husband, would Mrs Rosy be happy or sad?' We also included another type of counterfactual task: false

syllogisms. Unlike the counterfactual conditionals these do not involve a series of events followed by a counterfactual question. Rather the child has to imagine as false something that she knows from general knowledge is true. For example, imagine 'All sheep are purple' and that 'Jenny is a sheep', 'What colour is Jenny?' We found that performance on three counterfactual tasks (Riggs et al.'s location change, German & Nichols' short causal chains (but not the long causal chains), and the false syllogisms) were related, even after we took account of children's chronological age and language ability. This suggests that they share a common element.

We also measured children's inhibitory control, that is, their ability to resist acting on irrelevant information. Counterfactual thinking should make inhibitory demands because one has to resist giving the most obvious realist answer to the counterfactual questions and instead consider the alternative possible world. We used two popular measures of 3- to 4-year-olds' inhibitory control. In the Bear/Dragon task (see e.g. Kochanska, Murray, Jacques, Koenig, & Vandegeest, 1996) the child is introduced to two puppets: the nice bear and the naughty dragon. Each gives simple instructions such as 'touch your nose', which the child must carry out if they come from the bear, but ignore if they come from the dragon. Children in our sample (as is typical) found it easy to respond to the bear but sometimes mistakenly act on the dragon's instructions. In the Black/White Stroop task children see a series of cards and have to say 'black' to the white cards, and 'white' to the black cards. The typical error is to say the actual colour of the card. We found that the better children performed on our inhibitory tasks (resisting the dragon's instructions and saying the opposite colours) the more likely they were to give the correct answer to the counterfactual questions. This supports the idea that what is difficult for 3-year-olds when they try to think counterfactually is resisting knowledge of what is known to be true: either knowledge of current reality that she has just seen acted out, or illustrated with pictures, or a general fact she knows about the world, e.g. that all sheep are white. The counterpart to ignoring what is known to be true is also likely to be an important aspect of children's counterfactual thinking. Children also have to entertain as true things they know to be false. What develops at around the age of 4 is an ability to reason with false content, while ignoring the truth.

3 Counterfactual Possibilities

For many researchers this is the end of the developmental story. At 4 years (if not before) children are able to answer explicit questions about counterfactual worlds—prompted both by counterfactual antecedents (e.g. Riggs et al. 1998) and counterfactual consequents (Guajardo & Turley-Ames, 2004). However, there is good evidence that at least two further developments occur.

In her research with adult participants, Byrne argues that counterfactuals are represented as 'dual possibilities', i.e. that people hold in mind both the counterfactual possibility and the actual possibility (e.g. Byrne, 2005; Thompson & Byrne, 2002). More recently evidence from a reaction time study showed that when people read

counterfactual conditionals (if *p had been the case,* then *q would have occurred*) their reactions times were equally quick when they next read *p, q* or *not p, not q*. When participants read indicative conditionals (if *p*, then *q*), they were slower to read *not p, not q* than *p*, *q*, which suggests that the former was not explicitly represented (Santamaria, Espino, & Byrne, 2005).

Although evidence suggests that adults think of counterfactuals as dual possibilities, there is no evidence that the 4-year-olds who answer counterfactual conditionals, such as those used by Riggs et al. (1998) are thinking about two possibilities (see Beck et al. 2006). Perner and Rafetseder (this volume) argue that 4-year-olds who answer counterfactual conditionals correctly may only be using hypothetical reasoning. In other words, children, unlike adults, may not be constraining their thinking based on the real world. Several studies have explored later developments in children's counterfactual thinking. To some extent, they all question when children begin to represent both the real and counterfactual worlds—which we might think of as Byrne's criterion that they are dual possibilities.

Beck et al. (2006) ran a study with 3- to 6-year-olds comparing responses to counterfactual conditional questions (which they called standard counterfactuals) and a new counterfactual question, the open counterfactual. They argued that the open counterfactual required the child to think about the past point in time when either of two events could have happened: the one that became reality and the one that is now counterfactual. The open counterfactual question took the form, 'Could something else have happened?' and Beck et al. argued that it involved thinking about both the real and counterfactual possibilities. They also used a new undetermined possibilities task, which required children to acknowledge that at any one point in time multiple future outcomes are possible. In the game children saw a toy mouse run down a slide that split in two halfway down. Once the mouse was at the bottom of the slide children were asked the counterfactual questions, either 'What if he had gone the other way, where would he be?' (standard counterfactual conditional) or 'Could he have gone anywhere else?' (the new open counterfactual). On undetermined trials the mouse waited at the top of the slide and children were asked to put out mats to ensure that he was caught at the bottom. The correct cautious response was to put out two mats, implicitly acknowledging that the mouse could go either way. The common incorrect response was to place only one mat.

Children found the counterfactual conditional questions significantly easier to answer than the open counterfactuals. What's more, they found it relatively difficult to use two mats to cover the two possible outcomes in the undetermined trials. For both younger (3- and 4-year-olds) and older (5- and 6-year-olds), performance on open counterfactuals and undetermined trials were not different to each other, but both were significantly worse than performance on the standard trials. Beck et al. argued from this that when children answer standard counterfactual conditionals they are not thinking of them as possibilities that could have happened. Instead, this understanding is demonstrated in the open counterfactual trials, when children recog-

nized that the mouse could have gone the other way. Understanding of multiple possibilities is also demonstrated when the older children put out two mats in the undetermined trials. There has been little developmental research on understanding counterfactuals as possibilities, but the delay between standard conditionals and open counterfactuals has been replicated in another study (Beck & Crilly, 2009). One of the most interesting questions is which cognitive processes underpin children's thinking about counterfactuals and actual events as possibilities. This has yet to be tested, but one likely candidate is working memory. Answering a standard counterfactual conditional requires the child to focus on only the counterfactual world, while trying to ignore anything from the real world (as we discussed above). In contrast, thinking about counterfactual possibilities involves holding in mind both possible events simultaneously.

This prediction is particularly interesting because several authors have already suggested there may be a working memory component in counterfactual thinking. However, they have typically been preoccupied with the shifts around 3-to-4 years of age. Robinson & Beck (2000) and Guajardo & Turley-Ames (2004) both speculated that thinking about a counterfactual event may make increased working memory demands. Indeed, Robinson & Beck (2000) found evidence that performance on one working memory task predicted success on counterfactual conditionals. However, they did not control for language ability. When the same comparison was made in our recent battery study (Beck, Riggs, & Gorniak, 2009) we found a relation between counterfactuals and working memory, but this no longer held when we controlled for language ability. If children only start to hold two worlds in mind simultaneously when they can answer open counterfactuals, then it is not until this point that we should expect a role for working memory. Future research should investigate this possible relationship.

4 Comparisons between Worlds

Further evidence for developments in counterfactual thinking after the age of 5 comes from the literature on counterfactual emotions such as regret and relief. These emotional experiences rely not only on holding possibilities in mind, but also on making a comparison between reality and what could have happened. For example, if you are given a choice between two envelopes and the one you choose contains £10 then you will be pleased. However, you would likely re-evaluate your winnings if you learned that the unchosen envelope contained £100. This comparison between the actual and counterfactual world suggests that understanding counterfactual emotions is likely to be more cognitively demanding that thinking about counterfactuals as possibilities.

There is very little research on the development of counterfactual emotions. In one study, by Guttentag & Ferrell (2004), two people experienced the same negative outcome—for example, taking the left path round a pond to school, but then cycling into a fallen tree and falling off their bikes. One person always took that route, so for him there was no salient counterfactual alternative. However, the second person

normally took the right path, but on this occasion decided to take the left one. So for him there was a salient counterfactual in that he could have taken his normal route. 7-year-olds and adults rated the second person as more likely to feel regret for their choice of route (because there is a clear counterfactual with a better outcome with which to make a comparison). However, using stories in which the counterfactual outcome was worse than the actual outcome, Guttentag & Ferrell found that 7-year-olds did not understand why people might feel relief. Why these authors found a discrepancy between regret and relief is unclear, although one possibility is that both adults and children are more likely to engage in counterfactual thinking when they encounter a negative, rather than positive outcome (Roese, 1997; German, 1999).

One study has compared directly children's performance on open counterfactuals and regret using Guttentag & Ferrell's stories (Beck & Crilly, 2009). In the new counterfactual task a man drove his car down a road and had a choice of two routes. One road led to the shops, the other to the swimming pool. Once he had driven to the swimming pool, children could be asked a standard counterfactual, 'What if he had driven the other way, where would he be?' or an open counterfactual, 'Could he have gone anywhere else?' Thirty 5-and 6-year-olds performed poorly on the regret stories but significantly better on the open counterfactual questions. Counterfactual emotions require more than just recognizing that the counterfactual world was once a possibility that could have replaced the actual world. They also require children to take an additional step of directly comparing the emotional outcomes in the counterfactual and actual worlds. In keeping with this idea that there is an extra step involved, success on counterfactual emotion tasks lags behind success on tasks that only require thinking about counterfactuals as possibilities.

Another approach has been taken by Amsel and colleagues, who used a task where children experience regret or relief themselves, rather than judging it in others (Amsel & Smalley, 2000). Children made a choice between two cards, which led to different rewards. Children rated how they felt about the actual outcome of their choice, both before and after they learnt what the unchosen reward was. Using this procedure Amsel & Smalley found that although 5-year-olds could describe the counterfactual outcome (their feelings if they had picked the other card) they did not show any evidence of counterfactual emotions. Amsel & Smalley's task involved both regret and relief trials. More recently, Weisberg & Beck (2010) used a very simple version of this task in which children chose between one of two boxes and won a number of stickers. They found that children as young as 5 showed some evidence of regret, but relief was not seen until the age of 7. When children watched another person play the game they did not attribute counterfactual emotions. In line with Guttentag & Ferrell (2004) children up to 7 years old did not seem to understand counterfactual emotions in others.

What domain-general process might underpin children's thinking about counterfactual emotions. Guttentag & Ferrell (2004) argued that what is critical in thinking about counterfactual emotions is the comparison between the actual and counterfactual world. This goes beyond holding both in mind (which we speculated in the

previous section may be related to working memory demands) and seems likely to involve switching flexibly between the two worlds. Attentional flexibility undergoes substantial improvements in middle childhood and may well be critical in children's manipulation of the different information involved in understanding counterfactual emotions. In a recent battery study (Burns, Riggs, & Beck, submitted) we found evidence to support this proposal. A sample of one hundred and three 4- to 7-year-olds completed counterfactual and executive tasks. The key tasks of interest were the boxes task used by Weisberg & Beck (2010), which measured children's own experience of regret (compared to understanding why others feel regret), and a computerized switching task devised by Davisdson, Amso, Cruess Anderson, & Diamond (2006). In this task children watched as arrows appeared on either the left or the right side of a computer screen. The arrows pointed either straight down or diagonally (to the other side of the screen) and the child's job was to respond by pressing the left or right key. A measure of attentional switching based on the cost of changing from responding to a downward pointing arrow to a diagonal one, or vice versa, predicted whether or not children experienced regret on the boxes task. It is of note that none of our executive measures correlated with performance on a story task adapted from Guttentag & Ferrell (2004). It may be that understanding why other people experience counterfactual emotions makes greater demands on conceptual understanding (reflecting about other people's emotions and their decision-making) than on executive measures, or it may have been that as performance on the stories was quite poor, there was not enough variation to identify statistical relationships. Despite this, we conclude that there is evidence that experiencing regret, that is making a comparison between the real and counterfactual worlds, seems to be dependent on domain-general developments in attentional switching.

5 Conclusions

Our aim in this paper has been to propose that there are at least four distinct developments in counterfactual thinking. Each builds on the one that developed before, resulting in an increasingly sophisticated ability to think about alternative possible worlds. First, children are able to speculate about alternative worlds, but only if they do not contradict what they know to be true. At around 4 years children are able to play with the truth and falsity of possible worlds: they can speculate about things that they know are false and their ability to resist interference from the real world improves. Only after this do children appreciate that the special nature of counterfactual possibilities is that they could once have replaced the actual event—at this stage thinking counterfactually begins to involve simultaneously representing both the counterfactual and actual worlds. Finally, it is not until middle childhood, around 6- or 7-years-old, that children appear to make spontaneous comparisons, switching between reality and what might have been. Understanding how counterfactual thinking is related to causality may not be so straightforward if, as we argue, there is not one key development. However, there is potential for a much richer description of how children's

speculative thinking might underpin different aspects of understanding causality if we accept these distinct developments.

References

Andrews-Espy, K. (2004) 'Using developmental, cognitive, and neuroscience approaches to understand executive control in young children', *Developmental Neuropsychology* 26: 379–84.

Amsel, E. & Smalley, D.J. (2000) 'Beyond really and truly: children's counterfactual thinking about pretend and possible worlds'. In: *Children's Reasoning and the Mind*, P. Mitchell and K.J. Riggs (eds) pp. 121–47. Hove, East Sussex: Psychology Press.

Beck, S.R. & Crilly, M. (2009) 'Is understanding regret dependent on developments in counterfactual thinking?', *British Journal of Developmental Psychology* 27: 505–10.

—— & Guthrie, C. (in press) 'Almost thinking counterfactually: children's understanding of close counterfactuals' *Child Development*.

—— Riggs, K.J., & Gorniak, S.L. (2009) 'Relating developments in children's counterfactual thinking and executive functions', *Thinking and Reasoning* 15: 337–54.

—— —— —— (2010) 'The effect of causal chain length on counterfactual conditional reasoning', *British Journal of Developmental Psychology* 28: 505–21.

—— Robinson, E.J., Carroll, D.J., & Apperly, I.A. (2006) 'Children's thinking about counterfactuals and future hypotheticals as possibilities', *Child Development* 77(2): 413–26.

Brocki, K.C. & Bohlin, G. (2004) 'Executive functions in children aged 6 to 13: A dimensional and developmental study', *Developmental Neuropsychology* 26(2): 571–93.

Burns, P., Riggs, K.J., & Beck, S.R. (submitted) 'Executive control and the experience of regret'.

Byrne, R.M.J. (2005) *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA.: MIT Press.

—— (2007) 'Author's response: The rational imagination and other possibilities', *Behavioural and Brain Sciences* 30: 470–80.

Chan, A. & Hahn, B. (2007) 'Causal order effect in three- and four-year-olds' counterfactual reasoning'. Poster presented to the Biennial Meeting of the Society for Research in Child Development, Boston, USA.

Davidson, M.C., Amso, D., Cruess Anderson, L., & Diamond, A. (2006) 'Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching', *Neuropsychologia* 44: 2037–78.

Gergely, G., Bekkering, H., & Király, I. (2002) 'Rational imitation in preverbal infants', *Nature* 415: 755.

German, T.P. (1999) 'Children's causal reasoning: counterfactual thinking occurs for "negative" outcomes only', *Developmental Science* 2: 442–47.

—— & Nichols, S. (2003) 'Children's counterfactual inferences about long and short causal chains', *Developmental Science* 6: 514–23.

Guajardo, N.R. & Turley-Ames, K.J. (2004) 'Preschoolers' generation of different types of counterfactual statements and theory of mind understanding', *Cognitive Development* 19: 53–80.

Guttentag, R. & Ferrell, J. (2004) 'Reality compared with its alternatives: Age differences in judgments of regret and relief', *Developmental Psychology* 40: 764–75.

Handley, S.J., Capon, A., Beveridge, M., Dennis, I., & Evans, J.St.B.T. (2004) 'Working memory and inhibitory control in the development of children's reasoning', *Thinking and Reasoning* 10: 175–96.

Harris, P.L. (1997) 'On realizing what might have happened instead', *Polish Quarterly of Developmental Psychology* 3: 161–76.

—— (2000) *The Work of the Imagination*. Oxford, UK: Blackwell.

—— German, T.P., & Mills, P. (1996) 'Children's use of counterfactual thinking in causal reasoning', *Cognition* 61: 233–59.

Johnson-Laird, P.N. & Byrne, R.M.J. (1991) *Deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Kochanska, G., Murray, K., Jacques, T.Y., Koenig, A.L., & Vandegeest, K.A. (1996) 'Inhibitory control in young children and its role in emerging internalization', *Child Development* 67: 490–507.

Kuczaj, S.A. & Daly, M.J. (1979) 'Development of hypothetical reference in the speech of young children', *Journal of Child Language* 6(3): 563–79.

Lehto, J.E., Juujarvi, P., Kooistra, L., & Pulkkinen, L. (2003) 'Dimensions of executive functioning: Evidence from children', *British Journal Of Developmental Psychology* 21: 59–80.

Perner, J. (2000) 'About + Belief + Counterfactual'. In: *Children's Reasoning and the Mind*, P. Mitchell & K.J. Riggs (eds), pp. 367–401. Hove, East Sussex: Psychology Press.

—— Sprung, M., & Steinkogler, B. (2004) 'Counterfactual conditionals and false belief: a developmental dissociation', *Cognitive Development* 19: 179–201.

Peterson, D.M. & Riggs, K.J. (1999) 'Adaptive modelling and mindreading', *Mind and Language* 14: 80–112.

Riggs, K.J., Peterson, D.M., Robinson, E.J., & Mitchell, P. (1998) 'Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality?', *Cognitive Development* 13: 73–90.

Robinson, E.J. & Beck, S. (2000) 'What is difficult about counterfactual reasoning?'. In: *Children's Reasoning and the Mind*, P. Mitchell & K.J. Riggs (eds), pp. 101–19. Hove: Psychology Press.

Roese, N.J. (1997) 'Counterfactual thinking', *Psychological Bulletin* 121: 133–48.

Santamaria, C., Espino, O., & Byrne, R.M.J. (2005) 'Counterfactual and semifactual possibilities prime alternative possibilities', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(5): 1149–54.

Simoneau, M. & Markovits, H. (2003) 'Reasoning with premises that are not empirically true: Evidence for the role of inhibition and retrieval', *Developmental Psychology* 39(6): 964–75.

Thompson, V.A. & Byrne, R.M. (2002) 'Reasoning counterfactually: Making inferences about things that didn't happen', *Journal of Experimental Psychology*: *Learning, Memory, and Cognition* 28(6): 1154–70.

Weisberg, D.P. & Beck, S.R. (2010) 'Differences in children's understanding of their own and others' regret and relief', *Journal of Experimental Child Psychology* 106: 184–91.

6

Domain-Specific Causal Knowledge and Children's Reasoning about Possibility

David M. Sobel[1]

Before children enter elementary school, they appear capable of making predictions about causal relations in the physical (e.g. Bullock, Gelman, & Baillargeon, 1982), psychological (e.g. Wellman, 1990), and biological (e.g. Inagaki & Hatano, 1993) domains. Young children can also generate explanations of observed events across these domains as well (e.g. Schult & Wellman, 1997; Shultz, 1982; Wellman & Liu, 2007). Some have argued that these predictive and explanatory abilities indicate children possess multiple domain-specific mechanisms of causal reasoning, which develop independently (e.g. Gopnik & Meltzoff, 1997; Wellman & Gelman, 1998).

Young children also appear capable of engaging in various kinds of counterfactual inference (e.g. Beck, Robinson, Carroll, & Apperly, 2006; German & Nichols, 2003; Harris, German & Mills, 1996; Kavanaugh & Harris, 1999; as well as various chapters in this volume). These abilities also undergo particular developmental changes. These differences are often explained in terms of the development of more domain-general mechanisms or reasoning processes (e.g. Beck et al. 2006; Beck, Riggs, & Burns, this volume; Riggs, Peterson, Robinson, & Mitchell, 1998; Woodward, this volume). Some have even argued that such development can subsume apparently domain-specific knowledge (e.g. Riggs et al. 1998, suggest that the development of counterfactual reasoning abilities accounts for the development of false belief).

[1] Acknowledgments: NSF (DLS-0518161) and NIH (NIMH F31-MH12047) supported the research described in this manuscript. I would like to thank Claire Cook, Alison Gopnik, Sarah Munro, Deena Skolnick-Weisberg, Josh Tenenbaum, and all the members of the 2007 workshop on Counterfactual and Causal Reasoning at the University of Warwick for helpful discussion in preparing this manuscript.

I would like to argue that in addition to the development of particular domain-general mechanisms that allow children to engage in kinds of causal and counterfactual inferences, there is another influence on their reasoning: their domain-specific causal knowledge. I will examine the relation between children's existing causal knowledge within a domain and their ability to make various kinds of inferences about possibility. I will explore this hypothesis across three kinds of inferences: (1) Children's use of a causal inferential mechanism based on Bayesian inference (which involves forming and reasoning about potential hypotheses); (2) Children's judgments about fictional and real entities; and (3) Children's ability to reason about counterfactuals. Each line of evidence will suggest that the way in which children represent their causal knowledge supports the kind of inference they can make.

1 Physical and Psychological Knowledge in Reasoning about the Possibilities in Interpreting Ambiguous Causal Evidence

In previous work, my colleagues and I (Sobel, Tenenbaum, & Gopnik, 2004) introduced 3- and 4-year-olds to a machine that lit up and played music when particular objects were placed on it. We explained that objects that activated the machine were called 'blickets' (and that the machine was a 'blicket machine'). We showed children a set of identical wooden blocks and that of the first twelve blocks placed on the machine, either two or ten activated it (between subjects, in the *rare* and *common* conditions respectively). Children were also shown that the machine activated if at least one object placed on it was a blicket. Children then saw a test trial, in which two new (but still perceptually identical) blocks (A and B) activated the machine together, followed by object A activating the machine by itself. We asked children whether A and B were each a blicket.

Object A is clearly a blicket, and all children categorized it as such. The more interesting question concerns the status of object B. If we make certain assumptions about the way the blicket detector works (which I will describe below), what's demonstrated in the test trial provides no evidence about the causal status of object B—object B activates the machine with another object, which independently has the power to activate the machine. One could argue that the probability that object B is a blicket is equivalent to the base rate of blickets in the world. If children recognize this possibility, their treatment of object B should depend on whether they observed that blickets were rare or common. This is a different pattern of response based on just recognizing associations among events—object B has only ever been associated positively with the machine's activation; on this view of the world, it should be a blicket regardless of the base rate. 4-year-olds responded that object B was a blicket ~80% of the time in the common condition, but only ~20% of the time in the rare condition (registering the base rate). 3-year-olds, in contrast, were insensitive to the training, and

stated that object B was a blicket ~80% of the time in both conditions. Using a similar procedure, adults respond in a qualitatively equivalent manner as 4-year-olds (Griffiths, Sobel, Tenenbaum, & Gopnik, in press).

These data have been used to suggest that 4-year-olds (and adults) possess a causal inference mechanism that is well-described by Bayesian inference, constrained by particular pieces of prior knowledge (Griffiths et al. in press; Griffiths & Tenenbaum, 2007; Sobel, 2009; Sobel et al. 2004; Tenenbaum & Griffiths, 2003). These researchers do not suggest that children are engaging in explicit Bayesian computations; rather, these investigations suggest that Bayesian inference is a computational-level description (cf. Marr, 1982) of children's causal inference: a description of how the child integrates the data they observe with the causal knowledge they possess.

One aspect of this description is the idea that children make inferences over a hypothesis space of possible causal models that could explain the data they observe. For example, in the Sobel et al. (2004) procedure, when children observe the test trial, there are four hypotheses potentially consistent with these data:

h_1) that neither object is a blicket
h_2) that only object A is a blicket
h_3) that only object B is a blicket
h_4) that both objects are blickets.

The data children observe are equally inconsistent with hypotheses h_1 and h_3 because object A has to be a blicket. The data, however, are equally consistent with the other two hypotheses (h_2 and h_4). Bayesian inference rationally describes how the prior probability of blickets affects children's inferences. I will not describe the full computation model in detail here, but refer the reader to Tenenbaum & Griffiths (2003) and Griffiths et al. (in press) for a short and more detailed version of this description respectively.

Here I will consider just the following questions: what knowledge is necessary to form this hypothesis space, and do children possess this knowledge? First, children must recognize the temporal relation between cause and effect; placing an object on the blicket detector makes it activate and the detector's activation should not cause the experimenter to place an object on it. Second, any object's location in space should be independent of any other object's locations in space; placing an object on the machine should not cause the experimenter to place any other object on the machine. Similarly, if a blicket is in one location, it does not mean an object in another location is also a blicket. Given research on infants' causal perception (e.g. Leslie & Keeble, 1987; Oakes & Cohen, 1990), and pre-schoolers' causal knowledge (e.g. Bullock et al. 1982; Sophian & Huber, 1984), it seems reasonable to assume that both 3- and 4-year-olds understand these principles. Such knowledge limits the hypothesis space to $h_{1 \ldots} h_4$.

But children also need to understand that there is a particular kind of causal relation between objects and machine. Sobel et al. (2004, following Tenenbaum & Griffiths, 2003) called this the *activation law*: children must recognize that there is some

mechanism that relates blickets to the detector's activation in a deterministic (or near deterministic) manner. Without this, the test trial is more consistent with object B being a blicket than the base rate—even though object A is a blicket, it might have failed to be efficacious when A and B were presented together. This is how 3-year-olds responded. I do not mean to suggest that 3-year-olds are indeterminists (this seems highly unlikely, and there is some direct evidence against this possibility, e.g. Kuzmak & Gelman, 1986; Sobel et al. 2009). Rather, 4-year-olds appear to interpret the causal relation between a blicket and the blicket detector as indicating there is a stable mechanism by which a blicket causes the detector to activate. It is doubtful that 4-year-olds know what the nature of that mechanism is (indeed, it is doubtful that adults know the nature of that mechanism, see Rosenblit & Keil, 2002). Rather, if children know there is something about a blicket that makes the machine activate, then they will treat the data in the test trial as ambiguous, and resort to the base rate. If they do not understand that such a mechanism must be present, they will be more likely to respond based on the associative information they observed.

There is some evidence that 4-year-olds assume causal relations are deterministic (e.g. Buchanan, Tenenbaum, & Sobel, 2010; Bullock et al. 1982; Schulz & Sommerville 2006). For example, Schulz and Sommerville (2006) demonstrated that 4-year-olds treat apparently stochastic data as indicating the presence of a hidden cause that worked deterministically; after observing stochastic evidence, children preferred to intervene on a novel event that had been hidden during the familiarization than a familiar event that worked stochastically. Further, several investigations suggest that preschoolers struggle with reasoning about various kinds of causal inferences that involve probabilistic contrasts (e.g. Kuzmak & Gelman, 1986; Perner, 1979; Sobel et al. 2009).

A more concrete way of examining what children know about the nature of the relation between objects and their causal efficacy is to investigate what children know about the mechanisms by which objects produce that efficacy. Do children know how a blicket can cause the machine to activate? To this end, I will turn to the study of children's understanding of the relation between objects' causal properties and stable, non-obvious properties of those objects.

Developing Knowledge of Objects' Non-obvious Properties. There is strong evidence that preschoolers relate objects' causal properties with their category membership (e.g. Gopnik & Sobel, 2000; Kemler-Nelson, 1995, Kemler-Nelson, Russell, Duke, & Jones, 2000), but there is also evidence that a robust understanding of this relation develops between ages 3–4 (e.g. Nazzi & Gopnik, 2000). There is also evidence that children believe that objects with shared category membership (indicated by a common label) will share internal properties, and even that 4-year-olds recognize that certain categories are defined by their insides while others are 'inside-irrelevant' (e.g. a jar; Gelman & Wellman, 1991). Again, however, there are some suggestions that a robust understanding of this relation develops between ages 3–4 (e.g. Gottfried & Gelman, 2005).

Are these two sets of findings related? Do children recognize that objects with shared causal properties share insides, and that those insides might be responsible for those

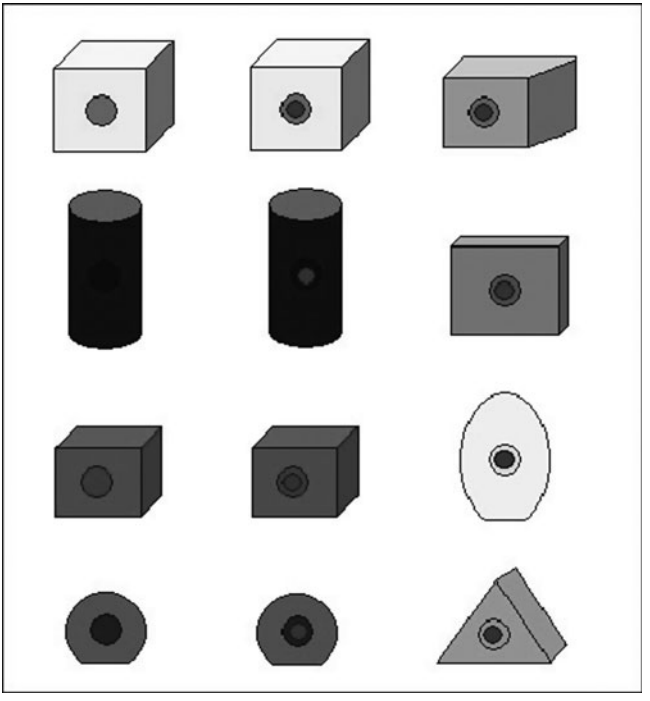


Figure 6.1 Schematic example of the stimuli used in Sobel, et al. (2007).

causal properties? Sobel, Yoachim, Gopnik, Meltzoff, & Blumenthal (2007) demonstrated that 4-year-olds, but not 3-year-olds, connect causes with insides. In one experiment, 4-year-olds transferred the causal efficacy of an object to another when its internal part was moved. In another, 4-year-olds were shown an object with both an internal and external part, and two identical objects, one with only the internal part and one with only the external part. When shown that the object with both parts had causal efficacy on a machine (it made it activate), children chose the object with the common inside as the one that would also make the machine activate.

Most importantly, 4-year-olds but not 3-year-olds relate causes and insides in light of competing perceptual information. Children were shown the blicket machine (although it was not labeled as such—it was only referred to only as a machine) and a set of objects like those shown in Figure 6.1: two identical objects and a unique one. Children observed that one of the two identical objects and the unique object activated the machine and other object did not. They were then shown that the member of the pair that activated the machine had an internal property and were asked which other object had the same inside. 4-year-olds responded based on the causal efficacy, and chose the unique object; 3-year-olds responded based on perceptual similarity.

What these data suggest is that 4-year-olds, but not younger children, understand that objects' causal and non-obvious (and in this case, internal) properties are related. Stronger evidence for this point come from some recent data collected in my lab. We told 3- and 4-year-olds that the blicket detector was a 'blicket machine' and 'things with blickets inside made the machine go'—providing them with strong information about how to relate the insides of the objects with the machine's efficacy. We then showed children that objects with internal parts (that were labeled 'blickets') activated the machine, and objects without internal parts did not activate the machine. We also showed children that the detector activated if at least one object with a blicket inside was on it. We then showed children two new objects (A and B), which activated the machine together. The door on object A was opened to reveal it was empty. Children had no trouble inferring whether each object contained an internal part (they claimed A did not, and B did). The critical question was an *intervention* question—we asked children to make the machine go. They had observed an action that was effective—placing both objects on the machine. But if children recognize that the internal part is responsible for the object's causal property, then they should recognize there is no need to put object A on the detector to make it activate. The majority of the 4-year-olds registered this information, and only placed object B on the machine. They did this significantly more often than the 3-year-olds, who were more likely to imitate what they had observed as efficacious (place A and B on the machine together).

Mr. Blicket and Blicket Machines. The data presented above suggest that children develop an understanding of the relation between causes and insides (and potentially that those insides are responsible for objects' causal properties) between the ages of 3 and 4. More generally, these data suggest that children develop an understanding of the nature of physical causal mechanisms between ages 3–4, consistent with a set of similar findings (e.g. Buchanan & Sobel, in press; Bullock et al. 1982; Gottfried & Gelman, 2005; Shultz, 1982). Is it possible that this developing knowledge allows children to recognize the inherent ambiguity in the rare-common manipulation described above? That is, if 3-year-olds could be made aware of the relation between objects' causal properties and another stable, non-obvious property, would they interpret the data they see as ambiguous and rely on the base rate information inherent in the procedure?

Sobel and Munro (2009) presented 3-year-olds with Mr. Blicket (see Figure 6.2): a blicket machine with cardboard eyes that spontaneously activated contingent on the experimenter as if they were having a conversation (akin to a manipulation used by Johnson, Slaughter, & Carey, 1998 on infants). After the experimenter and child conversed with Mr. Blicket, the experimenter told the child that Mr. Blicket would tell him/her what objects he likes. Drawing from the theory of mind literature, we chose to focus on Mr. Blicket's desires because by the age of 3, most children recognize that their own desires can be different from others' (Repacholi & Gopnik, 1997), that fulfilled and unfulfilled desires have causal consequences (Wellman & Woolley, 1990), and that shared preferences for objects relate to non-obvious properties of those objects (Fawcett & Markson, 2010). These data all suggest that 3-year-olds might understand that

Figure 6.2 Mr. Blicket.
Source: Photograph taken by the author.

Mr. Blicket wants particular objects for some reason, which is stable in the context of the experiment.[2]

Our first experiment with Mr. Blicket was an extension of the Sobel et al. (2007) procedure, in which 3-year-olds were shown the objects from Figure 6.1. For each set, they observed that Mr. Blicket liked one member of the identical pair and the unique object, but not the other member of the identical pair. They were then shown that the inside of the object in the pair that he liked had an internal property and were asked which other object had the same insides. The majority of 3-year-olds in this condition chose the unique object. To ensure that children were not just more engaged by this task than by a task involving a blicket machine, a control group of 3-year-olds was given the same procedure using a blicket machine that spontaneously activated at the beginning of the experiment, while the experimenter talked to the child. The machine's activation had a similar pattern, but it was not contingent on the experimenter. In this condition, 3-year-olds responded based on the external perceptual features of the objects.

Given the possibility that 3-year-olds recognized that Mr. Blicket's desires were related to a stable, non-obvious property of objects, we wondered if 3-year-olds would be sensitive to base rate information if the Sobel et al. (2004) procedure was replicated with Mr. Blicket's desires as opposed to a blicket machine. A new group of 3-year-olds

[2] Of course, an individual's desires can change over time. However, in this procedure, we were concerned with whether children recognized that if an agent liked two things at a particular moment in time, those two things might share some non-obvious feature (given that the external perceptual features of the objects did not offer a guide to Mr. Blicket's desire for those objects).

participated in this procedure: they were introduced to Mr. Blicket in the same manner, trained that he either liked few or many things (two or ten of the first twelve identical blocks in a box, the *rare* and *common* conditions respectively) and that he 'activated' if at least one thing that he liked was on him. They were then shown analogous data from the test trial: that he liked objects A and B together and he liked object A by itself. All children stated that he liked object A. Ninety-three per cent of 3-year-olds in the common condition stated he liked object B compared to the 44% who said so in the rare condition.

Three controls were critical. First, we wanted to ensure that children were not relying on the spontaneous activation of the detector, but rather understood that there was a difference between the causal relations indicated by Mr. Blicket's desire towards an object and an object making the blicket machine activate. Another group of 3-year-olds were introduced to a spontaneously activating blicket machine. They received a procedure analogous to the rare condition of this experiment, using that machine. These children categorized object B as having efficacy 72% of the time, significantly more often than children in the Mr. Blicket desire condition. This suggests that when the same data are presented as a machine activating, children respond associatively, and do not use the base rate information.

Second, we wanted to ensure that children did not simply find Mr. Blicket more interesting, and thus brought more cognitive resources to the task. Another group of 3-year-olds were introduced to Mr. Blicket in the same manner and given the identical procedure to those in the rare condition, except that they were told that Mr. Blicket's activation indicated that he was 'thinking about' an object. 3-year-olds have a much more difficult time understanding the causal structure of thought than of desire (e.g. Flavell, Green, & Flavell, 1995; Wellman, Cross, & Watson, 2001), and the data were consistent with these findings: again 72% of the children stated object B had efficacy, significantly more often than children in the desire condition.

Third, we also wanted to ensure that 3-year-olds were not confused when asked to relate Mr. Blicket's activation with his thoughts as opposed to his preferences. Desires tend to be based on fairly stable dispositions, while thoughts might be more fleeting and why an individual thinks about an event might not always be stable (e.g. Eisbach, 2004), particularly with regard to properties the objects might have. In a final experiment, we introduced a new group of 3-year-olds to Mr. Blicket (we called him Mr. B. in this experiment to avoid confusion), and told them that he knew which of the objects were 'blickets' and his activation indicated that a blicket was on him (i.e. he just acted like a blicket machine). Again, knowledge is a mental state that is not as well-understood by 3-year-olds as desire (e.g. Gopnik & Slaughter, 1991), so we predicted that children would not understand this particular activation law, and not use the base rate information. Indeed, they did not, categorizing object B as a blicket 80% of the time even though blickets were rare.

In each of these controls cases, our hypothesis was that children would not register the relation between a stable property of the objects and those objects' efficacy (i.e.

whether they activated the spontaneously activating machine or Mr. Blicket/Mr. B agent). However, a critical prediction of this account is that small subsets of 3-year-olds would register these relations. We performed a post-test on children in the rare conditions of the Mr. Blicket desires procedure, and all three of the control procedures (ninety children in all). After their participation in the procedure and a delay, all children received the same internal properties procedure, using the stimuli shown in Figure 6.1. Across the procedures, the higher the percentage of responses in which the child related the internal properties of the objects to their causal properties, the more likely children were to respond consistently with the base rate (and thus to have interpreted the data from the test trial as not revealing anything meaningful about the causal status of object B). This correlation was significant even when unrelated measures of general cognition were factored out. 3-year-olds who understood that the causal efficacy of the objects was related to a stable, non-obvious property were more likely to observe the ambiguity in these procedures, and potentially use the base rate information appropriately.

An important facet of these data is that across the 'rare' conditions, children were shown the same data; what differed was the way in which the experimenter described the data, which we assumed tapped into (or not) children's existing causal knowledge. Because 3-year-olds knew a great deal about the relation between desire and non-obvious properties, children were able to interpret the data they observed as ambiguous. In contrast, 3-year-olds do not appear to understand the relation between objects' properties and the working of a physical machine or why people think or know about certain things but not others, and as such failed to interpret the present data as ambiguous in the base rate experiment.

The final point worth making here is about the definition of 'domain'. Concepts like *desire, thought,* and *knowledge* are all part of the psychological domain, and seem ontologically different from the physical mechanisms involved in how a blicket detector might activate. I do not want to suggest that the domain-specific knowledge necessary to engage in the kind of hypothesis formation for Bayesian inference means that *any* psychological inference will emerge before *any* physical inference. Rather, what is critical to children being able to engage in this kind of causal reasoning is their understanding of particular causal mechanisms within a domain of knowledge.

2 Physical and Biological Knowledge in Judgments of Fantasy

The previous section suggested that 3-year-olds understand certain aspects of the causal structure in the psychological domain before they understand certain aspects of causal structure about the physical world. In this section, I want to examine children's general causal knowledge about the physical world compared to the biological world, focusing on a particular kind of judgment: about what is and is not possible. Judgments of

possibility are clearly based on experience—usually if we experience something, we believe it is possible. However, there are many objects and entities with which children might not have had direct experience that they still believe are possible. Similarly, numerous experiments in cognitive development present children with scenarios that actually are impossible (through various kinds of sleight of hand, for example, the 'shrinking machine', see e.g. DeLoache, Miller, & Rosengren, 1997). Because a child's causal knowledge makes a set of ontological predictions regarding what is and is not possible in the world, broad differences across these domains might interact with children's judgments about the possibility or impossibility of objects or events.

There are many studies of children's judgments of the fantasy/reality difference (e.g. Harris et al. 1991; Morison & Gardner, 1978; Samuels & Taylor, 1994; Shtulman & Carey, 2007; Subbotsky, 1994). Most of these suggest that children generally recognize that impossible events are impossible and real events are not. However, most of these studies do not consider judgments across domains of knowledge, or how children's developing causal knowledge in a domain might affect their judgments.

Biological vs Physical Knowledge. Infants understand various violations of causal structure in the physical environment (e.g. see reviews by Baillargeon, 2002; Spelke et al. 1992), and although paradigms that use action instead of looking time show a delay in children's understanding (e.g. Keen, 2003), there is little doubt that by age 4, children have a good understanding of certain aspects of physical causality. Further, the previous section detailed 4-year-olds' sophisticated causal reasoning abilities about certain kinds of physical events such as whether and what makes a machine activate (see Gopnik et al. 2001; Siegler 1976; Sobel et al. 2004; Sobel et al. 2007). Four-year-olds can also generate various mechanistic explanations for relations among physical events (Bullock et al. 1982; Shultz, 1982).

Children's biological knowledge, in contrast, is not as well developed. Children recognize differences between animate and inanimate objects, and assign psychological traits only to the former (e.g. Gelman & Opfer, 2002; Saxe, Tenenbaum, & Carey, 2005). But, most research on children's biological concepts proposes that biological knowledge is still developing during and after the pre-school years (e.g. Carey, 1985; Hickling & Gelman, 1995; Inagaki & Hatano, 1993, 2002; Rosengren et al. 1991; Springer, 1995). In general, these investigations suggest that during the preschool years, children appear to have a richer understanding of the physical than the biological domain.

A notable exception comes from a set of studies on children's explanations of impossible events. Schult & Wellman (1997; see also Wellman, Hickling, & Schult, 1997) investigated how preschoolers apply knowledge of folk physics, psychology, and biology to provide explanations of various types of human action. They found that 3- and 4-year-olds explained impossible physical events—such as a character wanting to float in the air—in terms of physical principles (e.g. gravity), and impossible biological events—such as a character wanting to hang from a tree branch forever—in terms of biological principles (e.g. limb fatigue). They concluded that young children have at least three basic

explanatory systems for providing explanations about physical, psychological, and biological events. Critically, Schult & Wellman (1997) asked children to explain why impossible events were impossible (i.e. could not happen). They found few domain differences between explanations of impossibility in the physical and biological domains.

These investigations suggest that young children have a relatively firm understanding of what is and is not possible across domains of knowledge. More importantly, these data suggest that children also understand the nature of the causal structures within these domains at around the same time. This is somewhat inconsistent with the conclusion from the previous section: that children's understanding of causal structure differs across domains and affects various inferences they make—in this case about the nature of possibility. At issue might be how children conceptualize *impossibility*. Shtulman & Carey (2007) found that 4-year-olds had little difficulty judging impossible and ordinary possible events as such, but also tended to judge improbable (but possible) events as impossible. They suggested that young children make judgments of impossibility based on whether they can 'imagine circumstances that would allow an event to occur' (2007: 1015). Such an ability potentially relies on children's experience with the world (or memory of that experience), but also should rely on children's developing causal knowledge. If this were the case, then we would expect to see differential judgments based on both of these pieces of information. Judgments about known possible events might be based on experience (children would presumably judge a microwave as real, even if they don't know how one works). In contrast, judgments about unknown or impossible events (including whether the event is impossible) might be based more on the child's existing causal knowledge.

The Ontology of Machines. Cook & Sobel (2011) examined this hypothesis by investigating children's judgments about whether various kinds of machines were possible. We were interested in machines because young children are clearly exposed to the category, but the increasing prevalence of technology suggests that judgments about the possibility of various kinds of machines reflect deep ontological commitments about what causal relations are possible.

We asked 4-year-olds, 6-year-olds, and adults about four categories of machines: (1) machines that exist in the world and we believed were familiar to children (e.g. a *radio*: a machine that plays music); (2) machines that exist in the world and would only be familiar if the child had a particular experience that introduced him/her to the machine (e.g. a *metal detector*: a machine that beeps if metal is nearby); (3) machines that were impossible because they violated an aspect of physical causality (e.g. a *shrinker*: a machine that makes a big crayon into a small crayon); and (4) machines that were impossible because they violated an aspect of biological causality (e.g. a *plant speaker*: a machine that makes flowers talk). Figure 6.3 shows the percentage of times children and adults judged each of these types of machines were real.

All three groups judged the familiar possible machines as real and both categories of impossible machines as make-believe more often than chance. However, 4-year-olds showed a difference in their categorization of the physical and biological impossible

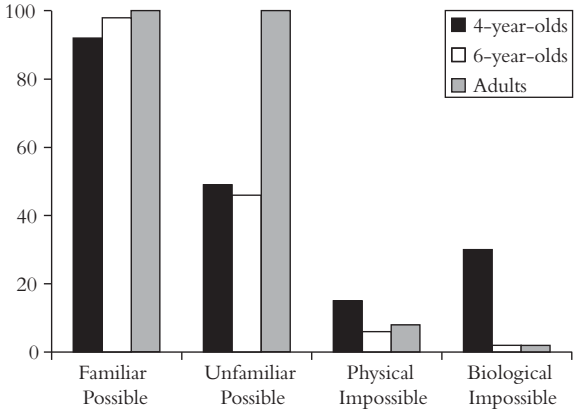


Figure 6.3 Percentage of machines in each category judged to be real (as opposed to impossible). Four-year-olds, but not 6-year-olds, judged impossible machines differently across the physical and biological domains.

Source: Data from Cook & Sobel (2011).

machines, saying the latter were more likely to be real than the former. Older children did not show this difference, possibly because their understanding of causal knowledge in the biological domain is more developed.

Both 4- and 6-year-olds responded at chance levels about unfamiliar machines. There are a number of possibilities for this pattern of performance. Because we could not standardize the experiences children had prior to coming into the lab, some children might have had experience with some of these items. If children recalled their experiences with the machine in question (e.g. metal detectors at an airport), they should say that it is real. If children did not recall the experience or never had the experience in the first place, they might treat the machine as Shtulman & Carey's (2007) improbable items, categorizing it as make-believe because they can't access a situation in which it is real.[3] It could also be possible that both age groups might lack specific pieces of causal knowledge necessary to recognize that the mechanisms for some of these machines are possible. The present data cannot discern between these possibilities. Finally, as children got older they tended to categorize more of the biologically impossible items as make-believe. Judgments in no other category significantly correlated with age. This correlation was present even just among the 4-year-old sample, suggesting that some biological knowledge develops during this time. In general, these data are problematic for the idea that experience alone constrains

[3] A detailed examination of these data found that the variance in the unfamiliar machine condition was greater than that of the familiar machine condition (and either impossible category as well). This suggests a greater level of individual differences of responses in this condition, consistent with the analysis described in the text.

children's fantasy judgments (clearly, children do not observe the physical or biological violations we asked them about in the real world). Rather, we suggest that children's causal knowledge constrains what they believe is impossible. This is consistent with the arguments from the previous section about Mr. Blicket. In that investigation, for a child to reason in a manner consistent with Bayesian inference (a domain-general judgment), they needed a particular piece of specific knowledge within a domain crucial to understanding something about the causal laws of that domain. Here, children need similar domain-specific knowledge to make a judgment about whether something is real or make-believe. They need to know that impossible entities are ones that violate a causal law.

Children's Fictional World Construction. The experiment described above suggests that children might use their understanding of causal structure to recognize that certain physical violations are make-believe, and that children are more credulous towards violations of biological structure. If this is the case, then we might expect children to treat violations of physical and biological causality differently in other tasks that require them to make judgments about fantasy worlds. To examine this, we focused on how children constructed fictional worlds in the context of writing a novel story.

Intuitively, constructing a fictional world involves more than just recognizing the difference between what is real and what is not. Skolnick & Bloom (2006a) point out that children must appreciate the causal structure of the fictional world in question, and what kinds of causal violations (if any) that world can support. In a set of empirical studies (Skolnick & Bloom, 2006b), they found that children do not treat all fiction as one possible world. Rather, they suggest that children appreciate particular boundaries among fictional worlds—for example, children believe Batman and Spongebob come from different fictional worlds, and that each believes the other is fictional. Batman, however, thinks Robin or the Joker is real.

This argument suggests that if children are asked to construct their own fictional world, then that construction will be relatively coherent in the kinds of causal laws that exist in the world. We (Sobel & Weisberg, in preparation) investigated this question by asking 4-year-olds to construct a novel story through a set of picture choices. Children were told that they were going to write a story about a novel character. Children saw six pairs of pictures. One picture in each pair depicted the character engaging in an ordinary event. The other picture showed the character engaging in an event with the same outcome, but in which the character violated some aspect of physical or biological causal structure. There were an equal number of physical and biological violations across the task. For example, in one picture pair, children could choose to make the protagonist live with his mom and dad, brother and sister, or make him live with his fifty-two mothers and fathers and sixty-seven brothers and sisters (an example of biological violation). In another picture pair, children could choose whether the character walked outside to play through an open door or through a wall (an example of a physical violation).

We coded three aspects of children's story construction: (1) How many fantasy choices children made; (2) How consistent each child was in his/her choices; (3) Whether responses differed between the physical and biological items. For the most part (79% of the time), children chose the ordinary pictures. Individual children were also relatively consistent in their choices. We considered the number of times children switched from an ordinary response to a fantasy response (or vice versa) over the course of the trial. For any given picture pair, the probability that children would switch their response type was only 23% (well under chance levels) and the majority of children (69%) made zero or one switch. Thus, once a child made an ordinary response, he/she was likely to continue making ordinary responses, and once a child made a violation response, he/she was likely to continue making responses of that type. When children did differ in their responses, they systematically did so across domains: they were more likely to have the character violate aspects of physical causality than biological causality, even though the overall level of such responses was rare.[4]

Are these data consistent with the previous findings about children's ontological commitments about machines? I suggested above that 4-year-olds recognized more impossible physical machines than biological machines as such because at this age, they understood the causal structure of the physical world more generally than the causal structure of the biological world. Here, when given the choice between constructing a story featuring an ordinary physical or biological event and a corresponding one that violates an aspect of causal structure in that domain, children made more physical violations than biological ones. If children misunderstand the biological violations as possible, they might be more inclined to choose biological violations more often than physical ones, which would be inconsistent with these data. However, this is not necessarily the case, as the ordinary options were clearly possible (and presumably would be recognized by children as possible). Thus children's failure to select the biological violations in the construction of a fictional world is not informative since there is always a possible alternative that could be seen as a safer choice to maintain coherence. Critically, it is important to remember that children are doing so in the context of generating a story—a *fictional* world. If children are generating such worlds coherently (as suggested by Skolnick & Bloom, 2006a, among others), then their causal knowledge should guide this construction. Overall, these data suggest that children's developing causal knowledge across domains influences their reasoning on the same kind of task.

[4] One other aspect of these data was interesting. We also contrasted two types of protagonists—a human character on Earth with an obvious alien character from another planet (Zoltron from the planet Zolnar, who was a purple carrot). Our goal was to suggest that one character might be more likely to violate real-world causal knowledge. However, children were completely insensitive to the difference between these conditions (and made mostly ordinary choices across both conditions). This finding, however, might indicate another way in which children's immature causal understanding manifests; at age 4, it would appear that children do not recognize that different contexts influence what is possible (consistent with Beck's research discussed in the section on counterfactuals below).

3 Counterfactual Reasoning across Domains[5]

The previous sections have suggested that children's causal knowledge influences their judgments about ambiguity and about fantasy, both in judging fantasy from reality but also in constructing novel pieces of fiction. This section will consider the relation between children's causal knowledge and their ability to engage in types of counterfactual inference. Harris and colleagues (Harris, German, & Mills, 1996; Harris, 2000; Kavanaugh & Harris, 1999) proposed that children's ability to learn new causal relations is not only related to their ability to engage in counterfactual thinking, but more strongly, following Mackie (1974), that counterfactual reasoning plays a primal role in determining whether a causal relation is present. On this view, children develop the ability to engage in counterfactual reasoning quite early on, use this ability to learn about causal relations in the world, and this ability is present across domains of knowledge.

This hypothesis has some difficulty, as both German (1999) and Guajardo & Turley-Ames (2004) found that young children differed in their ability to generate counterfactuals when explaining events, depending on the valence of the event (whether the outcome was positive or negative) or whether the counterfactual added or removed antecedent events. Other researchers have challenged the hypothesis that young children can engage in counterfactual inferences early in development more directly. For instance, Riggs et al. (1998) suggested that counterfactual inferences place certain demands on executive functioning, which might be difficult for young children. They told 3- and 4-year-olds a story about a character (Peter) who was sick in bed. Another character (Sally) went off to the drugstore to get him medicine. While Sally was away, Peter received a call to go to the post office to help put out a fire. Children were asked where Peter would have been had there not been a fire. They found that 3-year-olds in particular did not perform well on the counterfactual questions and suggest this is due to children having to inhibit where Peter actually was. In contrast, young children had less difficulty with a similar prediction about a future hypothetical (i.e. predicting what would happen given a hypothetical antecedent). Such an inference does not have the same executive demands; the counterfactual requires the child to inhibit the present reality in favor of alternate events, while the future hypothetical does not require this inhibition. Riggs et al. (1998) further argued that such executive functioning might mediate reasoning on a false belief task; they found a correlation between children's counterfactual reasoning abilities and success on a standard false belief measure (see also Perner, Sprung, & Steinkogler, 2004).

Beck et al. (2006) followed up this study by considering a slightly different hypothesis. They found that while 3–4-year-olds might have some difficulty with standard

[5] These ideas and some of the research presented in this section were developed in collaboration with Alison Gopnik.

future hypothetical or counterfactual questions, what they really struggle with is a more general concept of possibility (consistent with the conclusions made by Shtulman & Carey, 2007, described in the previous section). In their experiments, 3- and 4-year-olds could reason about events that could occur in the future or events that could have been different in the past had antecedent events been different. But at this age, children struggled when asked to make an open-ended response to a similar counterfactual question that required they recognize two different outcomes of an event as both possible (and plan accordingly for either). They suggested that preschoolers might make future hypothetical and counterfactual inferences, like the ones described by Riggs et al. (1998) above, 'by thinking about the counterfactual event in isolation, without recognizing its temporal relationship to the actual event' (1998: 423). They conclude that children develop a general understanding of possibility that underlies their ability to engage in counterfactual reasoning.

There is clear merit to this account. For example, it nicely explains a finding presented by German & Nichols (2003), who showed that preschoolers were better at making counterfactual inferences about a direct causal relation as opposed to a more indirect cause of an event. They argue that this difference results from the complexity of the inference and not from difficulties with counterfactual reasoning *per se*. Beck et al.'s (2006) account suggests that in longer causal chains, children must recognize that there are multiple possibilities that lead to the particular consequent, while in a short causal chain, children only have to consider the relation between the antecedent and consequent, regardless of any other possibilities. This appears to be a more parsimonious interpretation of these data[6] (Perner and Rafetseder, this volume, generate a similar analysis of these findings).

Consistent with the points made in the previous sections, the hypothesis that I would advocate here is a slight modification to Beck et al.'s (2006) argument: that children's existing causal knowledge might interact with their developing concepts of possibility. For instance, Beck, Riggs, & Gorniak (2010) argue that after age 4 children are still developing

> the understanding that the alternative world was once a possibility that could have happened instead of what really did happen (at around 5 or 6 years; Beck et al. 2006),

[6] Beck, Riggs, & Gorniak (2010) failed to replicate the German & Nichols (2003) findings using similar stories. An interesting difference in these findings is that German & Nichols tested their participants in Greek while Beck et al. tested children in English. Beck et al. believed that this should not have affected the results. However, an examination of Greek morphology concerning counterfactual utterances (Iatridou, 2000) suggests that Modern Greek has a rich morphological structure regarding counterfactual utterances (presumably richer than English). Children might be sensitive to that difference when learning their language, and thus children learning Greek might understand counterfactual questions earlier than their English-speaking counterparts. This is not to say that children with Greek as their native language can reason about counterfactuals earlier than children whose native language is English; rather they might understand the complicated morphology necessary to elicit adult-like responses about hypothetical questions at a younger age, because competency in their language requires sensitivity to this morphology. This is an open question, but this interpretation is consistent with Beck et al. (2006)'s hypothesis that what is developing is a broad understanding of possibility.

> and the understanding of the emotional consequences of considering these counterfactual worlds, such as regret or relief at around 7 years (Guttentag & Ferrell, 2004; see also Beck & Crilly, 2009). (pp. 14)

My reasoning for modifying Beck et al.'s (2006) argument starts from the minor discrepancy mentioned here. Why might children understand possibility at age 5–6, but not the emotional ramification of considering those counterfactual worlds until age 7? Perhaps this difference is not meaningful, but could this be because understanding the *causal* relations between outcomes and their emotional ramifications, particularly about emotions that require metacognitive awareness, does not develop until relatively late (around age 7, e.g. Berti, Garattoni, & Venturini, 2000; Harris, Olthof, Terwogt, & Hardman, 1987)? More generally, children's ability to make counterfactual inferences might depend on the specific causal knowledge they possess about the inference they are asked to make.

To illustrate this point, consider the following experiment by McCormack, Butterfill, Hoerl, & Burns (2009; McCormack, Frosch, & Burns, this volume). They presented children with similar test trials to those used by Sobel et al. (2004), but without the training. They were interested in whether children could answer counterfactual questions about whether the machine would activate. They found that 5-year-olds responded accurately to both predictive and counterfactual questions,[7] and suggest that, 'whatever account is provided of the processes and/or representations underpinning children's causal learning on the blicket detector task must be one that can explain children's ability to make not only causal judgments, but also counterfactual ones' (McCormack et al. 2009: 1573). Paralleling the work I presented above, I would suggest that if their experiment were replicated using Mr. Blicket and asking about his desires, younger children would also show the ability to answer correctly on counterfactual questions.

While I have not performed the experiment described in the previous paragraph, I have considered this question by looking at the kinds of counterfactuals questions used by Riggs et al. (1998) about different psychological states—whether someone will be happy or sad given a fulfilled or unfulfilled desire and whether someone will be surprised given their knowledge about an event. Understanding the emotional ramification of a fulfilled or unfulfilled desire is understood by almost all 3-year-olds (e.g. Wellman & Woolley, 1990), while understanding the causal relations between knowledge and surprise is typically not understood until age 5 or later (e.g. Hadwin & Perner, 1991).

I (Sobel, 2002) told 3- and 4-year-olds stories about characters (acted out with dolls), which ended with an antecedent and then a consequent event.[8] On one story, the

[7] Interestingly, they found that 4-year-olds did not respond in a similar manner, inconsistent with my previous work, and the work presented in the section above. They discuss some possible reasons for this difference, which are beyond the scope of the present discussion.

[8] This procedure was done as part of my doctoral thesis, and also considered other kinds of events. For purposes of space, however, I am omitting other aspects of this procedure. Feel free to contact me for more information.

antecedent event fulfilled a character's desire, which resulted in a consequent of the character being happy. Children were asked how the character would feel if the desire was unfulfilled. On the other story, the antecedent established that the character was happy but ignorant of a particular event, which would surprise him (e.g. another character jumping out of a closet), and was surprised when this happened. Children were asked how the character would feel if he was not ignorant of the event that surprised him (i.e. would he feel happy or surprised?). Children were relatively accurate at answering the counterfactual question about another's desire, and relatively inaccurate at answering the counterfactual question about surprise. Further, neither of these responses correlated with performance on a standard set of false belief tasks.

I am not arguing that all counterfactual reasoning emerges early on, or that children's capacity to reason about counterfactuals is the only mechanism (or even, necessarily, *a* mechanism) for causal learning. Rather, I suggest that children develop coherent representations of specific causal knowledge within a domain, which might support counterfactual reasoning abilities. Cases where we see children engaging in counterfactual reasoning early on might indicate that they possess an understanding of the causal structure of the situation. Cases where children might lack that causal knowledge (such as in the ramification relating events with complex emotions) might show a developmental lag from a general understanding of possibility. Children *are* developing domain-general mechanisms related to counterfactual inference. For example, children's understanding of the language used to ask counterfactual questions seems important (see footnote 6).

Like Beck et al. (2006), I also suspect that children's broader understanding of possibility is relevant to generating counterfactual inferences. In another set of experiments (Sobel, 2004), I told 3- and 4-year-olds stories in which a character wanted to perform an impossible task, tried to do so, but could not (because the event was impossible). Children were asked a more 'open-ended' question—what could the character do different—which is similar to the questions that Beck et al. (2006) described were difficult for young children. A small percentage of the time (~30%), children stated that the character could not do anything different because the task was impossible, and there were some differences across stories in the physical, biological, and psychological domains. I also asked the same children to generate explanations of impossible events in the same domains (some cases were violations of a similar causal law; in other cases the causal law was different from what was asked in the counterfactual question, but in the same domain of knowledge). Children explained impossible events appropriately more frequently than they correctly rejected counterfactuals as possible (~40–70% of the time across the domains, with some differences among them). But critically, the frequency with which that child generated correct explanations of why impossible events were impossible predicted how often they generated correct responses on the open-ended counterfactual question. The latter question is potentially more difficult for a pre-schooler; it has an executive demand that the former does not (i.e. from the child's perspective, why would the experimenter ask for an alternative possibility if there is not one?). These data are consistent with Beck et al.'s

(2006) account that such open-ended questions are difficult for young children, but they are also consistent with the possibility that regardless of that domain-general development, these data suggest that children's existing causal knowledge influenced their ability to generate such alternatives.[9]

An important challenge towards how we understand the development of counterfactual inference in children is to consider how such inferences might be related to causal knowledge. For instance, a nice empirical investigation would be to consider whether Beck et al.'s (2006) approach holds over various domains of knowledge. I suspect that children have difficulty with understanding what is possible, but also that some of this understanding will be mediated by their developing causal knowledge across domains.

4 Concluding Thoughts

Domain-general mechanisms are critical to cognitive development. Bayesian inference over causal graphical models has suggested a long line of computational and empirical studies (see e.g. Gopnik & Schulz, 2007). Imagination and pretense has suggested clear investigations of children's understanding of the fantasy/reality distinction (see e.g. Taylor, 1999). Theories of executive function also make predictions about how children engage in pretense and understanding others' pretending (e.g. Frye, 2000). Executive function is also clearly important to generating and reasoning about counterfactual inferences (e.g. Riggs et al. 1998), as might be children's pretense (Kavanaugh & Harris, 1999). In all cases, it seems critical to integrate these domain-general approaches with the idea that the development of domain-specific causal knowledge might affect these inferences. An open question, critical to this discussion, is where domain-specific causal knowledge comes from. Answering that question, and illuminating how this integration takes place, are clear goals for future research.

References

Baillargeon, R. (2002) 'The acquisition of physical knowledge in infancy: A summary in eight lessons'. In: *Handbook of Childhood Cognitive Development*, U. Goswami (ed.), pp. 47–83. Malden, MA: Blackwell.

Beck, S.R. & Crilly, M. (2009) 'Is understanding regret dependent on developments in counterfactual thinking?', *British Journal of Developmental Psychology* 27: 505–10.

——, Riggs, K.J., & Gorniak, S.L. (2010) 'The effect of causal chain length on counterfactual conditional reasoning', *British Journal of Developmental Psychology* 28: 505–21.

[9] Hitchcock (this volume) gives a particular good example of a view I would endorse, concerning counterfactual inferences about desks made of Jello. We can reason about such a counterfactual, but such a counterfactual's relation to causal inference is not clear. In order to reason about such a counterfactual, one must already possess some causal knowledge (e.g. about the structural integrity of gelatin). Seelau, Wells, & Windschitl (1995) make a similar point.

—— Robinson, E.J., Carroll, D.J., & Apperly, I. A. (2006) 'Children's thinking about counterfactuals and future hypotheticals as possibilities', *Child Development* 77: 413–26.

Berti, A.E., Garattoni, C., & Venturini, B.A. (2000) 'The understanding of sadness, guilt, and shame in 5-, 7-, and 9-year-old children', *Genetic, Social, and General Psychology Monographs* 126: 293–318.

Buchanan, D.W., & Sobel, D.M. (in press) 'Mechanism-based causal reasoning in young children'. *Child Development*.

—— Tenenbaum, J.B., & Sobel, D.M. (2010) 'Edge replacement and nonindependence in causation', *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*. Portland, OR., pp. 919–24.

Bullock, M., Gelman, R., & Baillargeon, R. (1982) 'The development of causal reasoning'. In: *The Developmental Psychology of Time*, W. J. Friedman (ed.), pp. 209–54. New York: Academic Press.

Carey, S. (1985) *Conceptual change in childhood*. Cambridge, MA: MIT Press/Bradford Books.

Cook, C. & Sobel, D.M. (2011) 'Children's understanding of the fantasy/reality status of hypothetical machines', *Developmental Science* 14: 1–8.

DeLoache, J.D., Miller, K.F., & Rosengren, K.S. (1997) 'The credible shrinking room: Very young children's performance with symbolic and non-symbolic relations', *Psychological Science* 8: 308–13.

Eisbach, A. O. (2004) 'Children's developing awareness of the diversity in people's trains of thought', *Child Development* 75: 1694–707.

Fawcett, C., & Markson, L. (2010) 'Children reason about shared preferences'. *Developmental Psychology* 46: 299–309.

—— —— (in press) 'Children reason about shared preferences. *Developmental Psychology*'.

Flavell, J.H., Green, F.L., & Flavell, E.R. (1995) *Young Children's Knowledge about Thinking. Monographs of the Society for Research in Child Development* 60 (1, Series No. 243).

Frye, D. (2000) 'Theory of mind, domain specificity, and reasoning'. In: *Children's Reasoning about the Mind*, P. Mitchell & K.J. Riggs (eds.), pp. 149–68. East Sussex, UK: Psychology Press.

Gelman, S.A. & Opfer, J. (2002) 'Development of the animate-inanimate distinction'. In: *Handbook of Childhood Cognitive Development*, U. Goswami (ed.), pp. 151–66. Malden, MA: Blackwell.

—— & Wellman, H.M. (1991) 'Insides and essence: Early understandings of the non-obvious', *Cognition* 38: 213–44.

German, T.P. (1999) 'Children's causal reasoning: Counterfactual thinking occurs for 'negative' outcomes only', *Developmental Science* 2: 442–7.

—— & Nichols, S. (2003) 'Children's counterfactual inferences about long and short causal chains', *Developmental Science* 6: 514–23.

Gopnik, A. & Meltzoff, A.N. (1997) *Words, Thoughts, and Theories*. Cambridge: MIT Press.

—— & Schulz, L. (2007) *Causal Learning: Psychology, Philosophy, and Computation*. New York: Oxford University Press.

—— & Slaughter, V. (1991) 'Young children's understanding of changes in their mental states', *Child Development* 62: 98–110.

—— & Sobel, D.M. (2000) 'Detecting blickets: How young children use information about causal properties in categorization and induction', *Child Development* 71: 1205–22.

Gopnik, A., Sobel, D.M., Schulz, L., & Glymour, C. (2001) 'Causal learning mechanisms in very young children: Two, three, and four-year-olds infer causal relations from patterns of variation and co-variation', *Developmental Psychology* 37: 620–29.

Gottfried, G.M. & Gelman, S.A. (2005) 'Developing domain-specific causal-explanatory frameworks: The role of insides and immanence', *Cognitive Development* 20: 137–58.

Griffiths, T.L., Sobel, D.M., Tenenbaum, J.B., & Gopnik, A. (in press) 'Bayesian reasoning in adults' and children's causal inferences'. *Cognitive Science*.

—— & Tenenbaum, J.B. (2007) 'Two proposals for causal grammars'. In: *Causal Learning: Psychology, Philosophy, and Computation*, A. Gopnik & L. Schulz (eds), pp. 323–45. New York: Oxford University Press.

Guajardo, N.R. & Turley-Ames, K.J. (2004) 'Preschoolers' generation of different types of counterfactual statements and theory of mind understanding', *Cognitive Development* 19: 53—80.

Guttentag, R. & Ferrell, J. (2004) 'Reality compared with its alternatives: Age differences in judgments of regret and relief', *Developmental Psychology* 40: 764–75.

Hadwin, J. & Perner, J. (1991) 'Pleased and surprised: Children's cognitive theory of emotion', *British Journal of Developmental Psychology* 9: 215–34.

Harris, P.L. (2000) *The work of the imagination*. Malden, MA: Blackwell Publishers.

—— Brown, E., Marriott, C., Whittall, S., & Harmer, S. (1991) 'Monsters, ghost, and witches: testing the limits of the fantasy-reality distinction in young children', *British Journal of Developmental Psychology* 9: 105–23.

—— German, T., & Mills, P. (1996) 'Children's use of counterfactual thinking in causal reasoning', *Cognition* 61: 233–59.

—— Olthof, T., Terwogt, M.M., & Hardman, C.E. (1987) 'Children's knowledge of the situations that provoke emotion', *International Journal of Behavioral Development* 10: 319–43.

Hickling, A.K. & Gelman, S.A. (1995) 'How does your garden grow? Early conceptualization of seeds and their place in the plant growth cycle', *Child Development* 66: 856–76.

Iatridou, S. (2000) 'The grammatical ingredients of counterfactuality', *Linguistic Inquiry* 31: 231–70.

Inagaki, K. & Hatano, G. (1993) 'Young children's understanding of the mind-body distinction', *Child Development* 64: 1534–49.

—— —— (2002) *Young Children's Naïve Thinking about the Biological World*. New York: Psychology Press.

Johnson, S., Slaughter, V., & Carey, S. (1998) 'Whose gaze will infants follow? The elicitation of gaze-following in 12-month-olds', *Developmental Science* 1: 233–8.

Kavanaugh, R.D., & Harris, P.L. (1999) 'Pretense and counterfactual thought in young children'. In: *Child Psychology: A Handbook of Contemporary Issues*, L. Balter & C.S. Tamis-LeMonda (eds), pp. 158–76. New York: Psychology Press.

Keen, R. (2003) 'Representation of objects and events: why do infants look so smart and toddlers look so dumb?', *Current Directions in Psychological Science* 12: 79–83.

Kemler-Nelson, D.G. (1995) 'Principle-based inferences in young children's categorization: Revisiting the impact of function on the naming of artifacts', *Cognitive Development* 10: 347–80.

—— Russell, R., Duke, N., & Jones, K. (2000) 'Two-year-olds will name artifacts by their functions', *Child Development* 71: 1271–88.

Kuzmak, S.D. & Gelman, R. (1986) 'Young children's understanding of random phenomena', *Child Development* 57: 559–66.

Leslie, A.M. & Keeble, S. (1987) 'Do six-month-old infants perceive causality?', *Cognition* 25: 265–88.

McCormack, T., Butterfill, S., Hoerl, C., & Burns, P. (2009) 'Cue competition effects and young children's causal and counterfactual inferences', *Developmental Psychology* 45: 1563–75.

Mackie, J.L. (1974) *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.

Marr, D. (1982) *Vision*. New York: Henry Holt & Company.

Morison, P. & Gardner, H. (1978) 'Dragons and dinosaurs: The child's capacity to differentiate fantasy from reality', *Child Development* 49: 642–48.

Nazzi, T. & Gopnik, A. (2000) 'A shift in children's use of perceptual and causal cues to categorization', *Developmental Science* 3: 389–96.

Oakes, L.M. & Cohen, L.B. (1990) 'Infant perception of a causal event', *Cognitive Development* 5: 193–207.

Perner, J. (1979) 'Discrepant results in experimental studies of young children's understanding of probability', *Child Development* 50: 1121–27.

—— Sprung, M., & Steinkogler, B. (2004) 'Counterfactual conditionals and false belief: A developmental dissociation', *Cognitive Development* 19: 179–201.

Repacholi, B.M. & Gopnik, A. (1997) 'Early reasoning about desires: Evidence from 14- and 18-month-olds', *Developmental Psychology* 33: 12–21.

Riggs, K.J., Peterson, D.M., Robinson, E.J., & Mitchell, P. (1998) 'Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality?', *Cognitive Development* 13: 73–90.

Rosenblit, L. & Keil, F. (2002) 'The misunderstood limits of folk science: An illusion of explanatory depth', *Cognitive Science* 26: 521–62.

Rosengren, K.S., Gelman, S.A., Kalish, C.W., & McCormick, M. (1991)'As time goes by: children's early understanding of growth in animals', *Child Development* 62: 1302–20.

Samuels, A. & Taylor, M. (1994) 'Children's ability to distinguish fantasy events from real-life events', *British Journal of Developmental Psychology* 12: 417–27.

Saxe, R., Tenenbaum, J.B., & Carey, S. (2005) 'Secret agents: inferences about hidden causes by 10- and 12-month-old infants', *Psychological Science* 16: 995–1001.

Schult, C.A. & Wellman, H.M. (1997) 'Explaining human movements and actions: Children's understanding of the limits of psychological explanation', *Cognition* 62: 291–324.

Schulz, L.E. & Sommerville, J. (2006) 'God does not play dice: Causal determinism and preschoolers' causal inferences', *Child Development* 77: 427–42.

Seelau, E.P., Seelau, S.M., Wells, G.L., & Windschitl, P.D. (1995) 'Counterfactual constraints'. In: *What Might Have Been: The Social Psychology of Counterfactual Thinking*, N.J. Roese & J.M. Olson (eds), pp. 57–79. Majwah, NJ: Erlbaum.

Shtulman, A. & Carey, S. (2007) 'Improbable or impossible? How children reason about the possibility of extraordinary events', *Child Development* 78: 1015–32.

Shultz, T.R. (1982) *Rules of causal attribution. Monographs of the Society for Research in Child Development* 47 (1, Serial No. 194).

Siegler, R.S. (1976) 'The effects of simple necessity and sufficiency relationships on children's causal inferences', *Child Development* 47: 1058–63.

Skolnick, D. & Bloom, P. (2006a) 'The intuitive cosmology of fictional worlds'. In: *The Architecture of the Imagination: New Essays on Pretense, Possibility, and Fiction*, S. Nichols (ed.), pp. 73–86. Oxford: Oxford University Press.

—— —— (2006b) 'What does Batman think about SpongeBob? Children's understanding of the fantasy/fantasy distinction', *Cognition* 101: B9–B18.

Sobel, D.M. (2002) *Examining the Coherence of Young Children's Understanding of Causality: Evidence from Inference, Explanation, and Counterfactual Reasoning*. Unpublished doctoral dissertation, University of California, Berkeley.

Sobel, D.M. (2004) 'Exploring the coherence of young children's explanatory abilities: Evidence from generating counterfactuals', *British Journal of Developmental Psychology* 22: 37–58.

—— (2009) 'Integrating top-down and bottom-up approaches to causal learning'. In: *Neoconstructivism: The New Science of Cognitive Development*, S. Johnson (ed.), pp. 159–79. New York: Oxford University Press.

—— & Munro, S.A. (2009) 'Domain generality and specificity in children's causal inferences about ambiguous data', *Developmental Psychology* 45: 511–24.

——, Sommerville, J.A., Travers, L.V., Blumenthal, E.J., & Stoddard (2009) 'The role of probability and intentionality in preschoolers' causal generalizations', *Journal of Cognition and Development* 10: 262–84.

—— Tenenbaum, J.B., & Gopnik, A. (2004) 'Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers', *Cognitive Science* 28: 303–33.

—— Yoachim, C.M., Gopnik, A., Meltzoff, A.N., & Blumenthal, E.J. (2007) 'The blicket within: Preschoolers' inferences about insides and causes', *Journal of Cognition and Development* 8: 159–82.

—— & Weisberg, D.S. (in preparation) *Tell me a story: how ontological boundaries develop in a story construction task*. Manuscript. Rutgers University.

Sophian, C. & Huber, A. (1984) 'Early developments in children's causal judgments', *Child Development* 55: 512–26.

Spelke, E.S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992) 'Origins of knowledge', *Psychological Review* 99: 605–32.

Springer, K. (1995) 'How a naïve theory is acquired through inference', *Child Development* 66: 547–58.

Subbotsky, E.V. (1994) 'Early rationality and magical thinking in preschoolers: space and time', *British Journal of Developmental Psychology* 12: 97–108.

Taylor, M. (1999) *Imaginary Companions and the Children Who Create Them*. New York: Oxford University Press.

Tenenbaum, J.B. & Griffiths, T.L. (2003) 'Theory-based causal inference', *Proceedings of the 14th Annual Conference on the Advances in Neural Information Processing Systems*.

Wellman, H.M. (1990) *The Child's Theory of Mind*. Cambridge, MA: Bradford Books/MIT Press.

—— Cross, D., & Watson, J.K. (2001) 'A meta-analysis of theory of mind: The truth about false belief', *Child Development* 72: 655–84.

—— & Gelman, S.A. (1998) 'Knowledge acquisition in foundational domains'. In: *Handbook of Child Psychology, 5th edition*, D. Kuhn & R. Siegler (eds), pp. 523–73. New York: Wiley.

—— Hickling, A.K., & Schult, C.A. (1997) 'Young children's psychological, physical, and biological explanations'. In: *The Core Domains of Thought: Children's Reasoning about Physical, Psychological, and Biological Phenomena*, H.M. Wellman & K. Inagaki (eds), pp. 7–25. San Francisco, CA: Jossey-Bass.

—— & Liu, D. (2007) 'Causal reasoning as informed by the early development of explanations'. In: *Causal Learning: Psychology, Philosophy, and Computation*, A. Gopnik & L.E. Schulz, pp. 261–79. New York: Oxford University Press.
—— & Woolley, J.D. (1990) 'From simple desires to ordinary beliefs: The early development of everyday psychology', *Cognition* 35: 245–75.

7

Mental Simulation and the Nexus of Causal and Counterfactual Explanation

David R. Mandel[1]

1 Introduction

Attempts to make sense of specific episodes in the past, especially when they entail consequential, surprising, or unwanted outcomes, tend to involve an interrelated set of causal and counterfactual questions that people may pose to themselves or to others: Why did it happen? How could it have happened? How might it have been prevented? And, so on. Given the transactional nature of such questions, the answers provided may be regarded as explanations (Keil, 2006). Such explanations have long been explained themselves in terms of the functional benefit of prediction and learning that they confer when they are accurate (Heider, 1958). However, such explanations, especially in cases involving harm, also underlie people's moral cognitions and 'prosecutorial mindsets' (Tetlock et al. 2007), serving as bases for addressing other related 'attributional' questions such as: Who is responsible? Who is to blame? What response—for instance, in terms of punishment or compensation—would be fair? And, so on.

For a few decades now, experimental psychologists have sought to understand the cognitive, motivational, and functional bases for such post-event querying. An important part of that endeavor has focused on elucidating the nature of the relationship between the various forms of causal and counterfactual thinking, which appear to give rise to the answers people provide to such queries. In this article, I examine the role of *mental simulation* (Kahneman & Tversky, 1982a)—the cognitive process whereby possibilities are brought to mind through mental construction—in causal and counter-

[1] I wish to thank Jim Woodward and the editors for their insightful comments on an earlier draft of this chapter.

factual explanations. I begin in Part 2 by discussing reasons for my emphasis on *explanation* as opposed to thinking or reasoning.

In Part 3, I trace the development of the mental simulation construct from Kahneman & Tversky's (1982a) seminal chapter on the simulation heuristic, noting how other psychologists have drawn on their notions of simulation and counterfactual thinking. My aim is Part 3 is largely twofold. Although Kahneman & Tversky's brief chapter on mental simulation was highly generative of subsequent research on counterfactual thinking, many of the ideas sketched, or simply alluded to, in the chapter have not been adequately discussed. Hence, one aim here is to reflect, and possibly expand, on some of those notions. For example, I explore some process-related issues pertaining to mental simulation that have not previously been discussed in the literature. My second objective is to critically examine how theorists, largely in social psychology, have drawn on the simulation heuristic notion to make claims about the nature of causal explanation. In doing so, I review psychological research on adults (for overviews of research on children, see in this volume: Beck, Riggs, & Burns; McCormack, Frosch, & Burns; Perner & Rafetseder; and Sobel) that has tested these notions.

In Part 4, I summarize an alternative 'judgment dissociation theory' of counterfactual and causal explanations that has emerged in later work, largely in response to the earlier notions discussed in Part 3. In this account (e.g. Mandel, 2003b, 2005a), although mental simulations play a role in both causal and counterfactual explanations, the focus of each type of explanation is different. Specifically, causal explanations tend to focus on antecedents that were sufficient under the circumstances to yield the actual event, whereas counterfactual explanations tend to focus on (the mutation of) antecedents that would have been sufficient to prevent the actual outcome and others like it from occurring. These different foci lead to predictable dissociations in explanatory content, which have been confirmed in recent experiments (e.g. Mandel, 2003; Mandel & Lehman, 1996). The chapter concludes with a discussion of the compatibility of these ideas with the kind of interventionist account that Woodward (this volume) seeks to advance.

To set the stage for the following discussion, it is important to point out, as the opening paragraph suggests, that I am mainly concerned here with explanation of tokens (i.e. particular cases) rather than of types (i.e. categories of cases). The studies I review, which were largely the result of the generative effect of Kahneman & Tversky's work on the simulation heuristic, tend to focus on people's explanations of negative past outcomes, such as why a particular protagonist died or how he could have been saved rather than what the most probable causes of death are or how life expectancy might generally be improved. Whereas causal and counterfactual reasoning about types focuses on ascertaining 'causal laws' (Cheng, 1993), causal reasoning about tokens may draw on knowledge about causal laws to answer attributional queries in ways that need not generalize to other cases, but that nevertheless constitute 'causal facts'. Woodward (this volume) makes a similar distinction, and applies his interventionist analysis to type rather than token causation. Towards the end of the chapter, I shall return to this issue in order to reflect on the compatibility of interventionism and judgment dissociation theory.

2 Why Explanation?

I use the term *explanation* rather than other terms such as *thinking* or *reasoning* in this chapter for two reasons. First, I believe that much of the emphasis on counterfactual and causal thinking about tokens, at least, functions to support explanation. Explanations, as noted earlier, are transactional (Keil, 2006), and subject to conversational norms (see, e,g. Grice, 1975; Hilton, 1990; Wilson & Sperber, 2004). Thus, explanations not only depend on the explainer's understanding of the topic, but also his or her assumptions or inferences regarding what the explainee may be seeking in a response. A good explanation for one explainee therefore may not be so for another, provided their epistemic states differ (e.g. Gärdenfors, 1988; Halpern & Pearl, 2005) or they seek different kinds of explanation (see also Woodward, this volume). For instance, harkening back to Aristotle's four senses of (*be*)*cause* (see Killeen, 2001), an explainer might give one individual seeking a mechanistic 'material cause' account of an event quite a different explanation than he or she would give to another individual seeking a functional 'final cause' explanation of the same event.

The transactional quality of explanation also leads to my second reason for focusing on explanation, and that is to better reflect the reality of the experimental context in which participants are asked to provide responses to questions posed by researchers. In studies I subsequently review, participants are usually asked to read a vignette about a chain of events that culminate in the story's outcome. Participants are then asked to indicate what caused the outcome and/or how the outcome might have been different 'if only . . . '. Thus, the participant in a psychological experiment faces many of the same challenges that any explainer would face.

The challenges, however, are in many ways much greater in the experimental context because the tasks imposed on the participant often violate conversational rules that would normally help explainers decide how to respond appropriately. For instance, in many everyday situations the reason why an explanation is sought may be fairly transparent and well indicated by the question itself. When it is not, the explainer can usually ask for clarification before formulating their response. In contrast, the experimental context often intentionally obscures such cues and denies cooperative opportunities for clarification so that the purpose of the experiment or the hypotheses being tested may remain hidden from the participant, and also so that all participants within a given experimental condition are treated in the same way. Moreover, given that the experimenter both provides participants with the relevant case information and then requests an explanation of the case from them, it may suggest to participants that they are being 'tested' in some manner (which of course they are). As Woodward (this volume) correctly observes, in many of the vignettes used in psychological studies the causal chain of events leading from the story's beginning to its ending are fairly complete. Thus, asking for an explanation may seem as odd as the answer would appear obvious. While I don't think the peculiarities of psychological research necessarily invalidate the exercise, it is important to bear in mind that the data produced by

participants are attempts at explanation that are not only constrained by 'causal thinking', but also by other forms of social, motivational, and cognitive factors that may have little, if anything, to do with causal reasoning per se.

Trabasso & Bartalone (2003) provide a good example of this. For years, it has been widely accepted that counterfactual explanations that 'undo' surprising outcomes tend to do so by mentally changing abnormal antecedents. This 'abnormality principle' traces back to influential papers in the psychological literature on counterfactual thinking—namely, Kahneman & Tversky's chapter on the simulation heuristic and Kahneman & Miller's (1986) norm theory. Trabasso & Bartalone, however, observed that abnormal events described in vignettes in experiments on counterfactual thinking tended to have more detailed explanations than normal events. This is unsurprising, since they were unusual. When the level of explanation was properly controlled, they found that counterfactual explanations no longer favored abnormal antecedents. Of course, their findings do not prove the unimportance of abnormality as a determinant of counterfactual availability, but the findings do illustrate the ease with which contextual features in experimental stimuli that influence participants' explanations can be misattributed to fundamental aspects of human cognition. It would be useful for experimenters and theorists to bear this in mind, and I would hope that a focus on explanation, with all that it entails, may be of some use in doing that. For instance, the vignette experiments described in Hitchcock (this volume) might be profitably examined in these terms.

3 Mental Simulation: Towards a Psychology of Counterfactual and Causal Explanation

In the psychological literature, sustained interest in understanding the relationship between counterfactual and causal thinking can be traced back to a brief, but influential, chapter by Kahneman & Tversky (1982a) entitled 'The Simulation Heuristic.' In it, the authors attempted to differentiate their earlier notion of the availability heuristic (Tversky & Kahneman, 1973) from the simulation heuristic. Whereas the availability heuristic involves making judgments on the basis of the ease of mental recall, the simulation heuristic involved doing so on the basis of the ease of mental construction.

Kahneman & Tversky (1982a) did not say much about what specifically characterizes a simulation, though it is clear from their discussion of the topic that they regarded mental simulation as closely linked to scenario-based thinking, or what they have in other work (Kahneman & Tversky, 1982b) referred to as the 'inside view', and which they distinguish from the 'outside view'—namely, thinking that relies on the aggregation of statistical information across multiples cases, and which they argue is more difficult for people to invoke in the service of judgment and decision-making. From their discussion, however, it would seem reasonable to infer that their notion of mental simulation was less restrictive than the manner in which representation is

depicted in mental models theory (Johnson-Laird & Byrne, 2002), which, as I discuss elsewhere (Mandel, 2008), mandates that the basic unit of mental representation be expressed in terms of possibilities depicted in rather abstract form. Mental simulations would appear much more compatible with the representation of scenes or stories (with a beginning, middle, and end) than with the mere representation of possibilities.

A central theme running through all of Kahneman & Tversky's program of research on heuristics and biases is that a person's experience of the ease of 'bringing to mind' is often used as a proxy for more formal bases of judgment (e.g. see Kahneman, Slovic, & Tversky, 1982). For instance, in judging the probability of an event class, one might be inclined to judge the probability as relatively low if it is difficult to recall exemplars of the class (via the availability heuristic) or if it is difficult to imagine ways in which that type of event might occur (via the simulation heuristic). These heuristics ought to provide useful approximations to accurate assessments if mental ease and mathematical probability are highly correlated. However, they will increasingly lead people astray in their assessments as that correlation wanes in magnitude. Or, as Dawes (1996) put it, for a counterfactual—and even one about a particular instance or token—to be regarded as normative or defensible it must be 'one based on a supportable statistical argument' (1996: 305).

Kahneman & Tversky (1982a; Kahneman & Varey, 1990) proposed that mental simulation played an important role in counterfactual judgments, especially those in which an event is judged to be close to having happened or having not happened. In such cases, they noted, people are prone to mentally undoing the past. Mental simulations of the past tend to restore expected outcomes by mutating unusual antecedents to more normal states and they seldom involve mutations that reduce the normality of aspects of the simulated episode. They referred to the former norm-restoring mutations as *downhill changes* and the latter norm-violating mutations as *uphill changes* to highlight the respective mental ease and effort with which these types of counterfactual simulations are generated. A number of other constraints on the content of mental simulations may be seen as examples of the abnormality principle. Some of these factors, such as closeness, are discussed by Hitchcock (this volume) and reviewed in depth elsewhere (e.g. Roese & Olson, 1995).

It is clear, even from Kahneman & Tversky's brief discussion of mental simulation, that they do not regard all mental simulation as counterfactual thinking. The earlier example of using mental simulation to estimate the likelihood of an event by gauging the ease with which one can conjure up scenarios in which the judged event might occur offers a case in point. There is no presumption in this example of a counterfactual comparison. Nor does mental simulation even have to be an example of hypothetical thinking since the representations brought to mind might be regarded as entirely veridical. In this regard, mental simulation seems to be conceptually closer to the notion of imagining, but with the constraint that the function of such imagining is to inform judgments of one kind or another, often by using the ease of construction as a proxy for what otherwise would be a more laborious reasoning exercise.

Kahneman & Tversky (1982a) also proposed that mental simulation could play a role in assessments of causality:

> To test whether event *A* caused event *B*, we may undo *A* in our mind, and observe whether *B* still occurs in the simulation. Simulation can also be used to test whether *A* markedly increased the propensity of *B*, perhaps even made *B* inevitable. (1982a: 202–3)

Clearly, their proposal was measured. For instance, they did not propose that causal assessments required mental simulations. Nor did they propose that the contents of such simulations necessarily bound individuals to their seeming implications through some form of intuitive logic. Thus, at least implicitly, they left open the possibility that an antecedent that, if mutated, would undo the outcome could still be dismissed as a cause (and certainly as *the* cause) of the outcome.

Later works influenced by their ideas were generally less measured in their assertions. For instance, Wells & Gavanski, (1989: 161) stated that 'an event *will* be judged as causal of an outcome to the extent that mutations to that event would undo the outcome' (italics added), suggesting that a successful case of undoing commits the antecedent to having a causal status. Obviously, there are many necessary conditions for certain effects that would nevertheless fail to be judged by most as causes. For instance, oxygen is necessary for fire. In all everyday circumstances where there was a fire, one could construct a counterfactual in which the fire is undone by negating the presence of oxygen. Yet, it is widely agreed that notwithstanding the 'undoing efficacy' of the antecedent, it would not be regarded as a cause of the fire in question, unless the presence of oxygen represented an abnormal condition in that instance (e.g. see Hart & Honoré, 1985; Hilton & Slugoski, 1986; Kahneman & Miller, 1986).

In other cases, antecedents that easily pass the undoing test would be too sensitive to other alterations of the focal episode to be regarded as causes (Woodward, 2006). For example, consider a case in which a friend gives you a concert ticket and you meet someone in the seat next to you who becomes your spouse and with whom you have a child. If the friend hadn't given you the ticket, the child wouldn't have been born. But few would say that the act of giving the ticket *caused* the child to be born. Other intriguing cases of counterfactual dependence that fail as suitable causal explanations are provided in Björnsson (2006).

Another variant of overstatement in this literature has been to assume that all counterfactual conditionals have causal implications. For example, Roese & Olson (1995: 11) state that '*all* counterfactual conditionals are causal assertions' and that 'counterfactuals, by virtue of the falsity of their antecedents, represent one class of conditional propositions that are *always* causal' (italics added). The authors go on to explain that 'the reason for this is that with its assertion of a false antecedent, the counterfactual sets up an inherent relation to a factual state of affairs' (1995: 11). This assertion, however, is easily shown to be false. Consider the following counterexamples: (1) 'If my name were John instead of David, it would be four letters long.' (2) 'If

I had a penny for every complaint of yours, I'd be a millionaire!' (3) 'If the freezing point had been reported on the Fahrenheit scale instead of the Celsius scale that was actually in use, the value would have been written as 32 °F instead of 0 °C.' In the first example, the counterfactual simply highlights a descriptive property of the speaker's counterfactual name. In the second example, the counterfactual offers the speaker a way of modulating the delivery of the intended criticism, and is an instance of what Tetlock & Belkin (1996) call *counterfactual morality tales*. In the last example, the counterfactual simply expresses a state of equivalence. Any of these examples suffices to show that counterfactual conditionals are not necessarily causal statements.

Process-related Issues

It is unclear from Kahneman & Tversky's (1982a) comments about the role of mental simulation in causal assessment how precisely they envisioned that such a process might work. Take the claim that to test whether event *A* caused event *B*, we may undo *A* in our mind, and observe whether *B* still occurs in the simulation. How would one apply such a test in practice? Most authors seem to have interpreted this type of counterfactual 'but-for' test of causality in discrete terms. That is, either the mutation of the antecedent succeeds in undoing the outcome (or effect) in question or else it does not. In the former case, it would be deemed a cause, while in the latter case it would not. For instance, Roese & Olson (1995: 13) state that 'whether or not a specific antecedent counterfactually undoes an outcome influences perceptions of that antecedent's causal impact'.

As noted earlier, however, the importance of *mental ease* as a basis for judgment in the 'heuristics and biases' framework suggests an alternative in which the goodness of a causal candidate is judged on the basis of the ease with which its negation leads to the undoing of the outcome of the episode in the relevant simulation. This type of process, which surprisingly seems to have been overlooked in the counterfactual thinking literature, would appear to offer a better fit to Kahneman & Tversky's ideas about the use of heuristics in judgment than a discrete 'either it undoes the outcome or it doesn't' assessment. Using the ease of mental simulation as a criterion for causal selection might also offer a way around a key problem faced by the counterfactual but-for test of causality; namely, as noted earlier, that it is too inclusive, yielding too many necessary conditions that pass the test (Hesslow, 1988; Lombard, 1990). As Hilton, McClure, & Slugoski (2005: 45) put it, 'This plethora of necessary conditions brings in its train the problem of causal selection, as normally we only mention one or two factors in a conversationally given explanation'. If mental ease were the basis for causal selection, then even if there were numerous antecedents that passed a counterfactual test of causality, an individual might select as 'the cause' from the set of viable candidates (i.e. those that undid the outcome in the simulation) the one that was easiest to bring to mind through mental construction. Or, perhaps more accurately, the simulation that most easily comes to mind as a way of undoing a focal outcome selects itself as a basis for causal understanding simply by virtue of its salience. To my

knowledge, this hypothesis (namely, that ease of mental simulation provides a basis for judging the goodness of a putative cause) has yet to be tested.

The proposal that mental simulation could be used to test whether *A* markedly increased the propensity of *B* also deserves comment. The idea is interesting because it suggests that mental simulation may play a role in assessments of the *sufficiency* of a putative cause to yield a particular effect. Although Kahneman & Tversky (1982a) did not explain how such a process might work, there are at least three possible routes of influence. The first would be to simulate a putative cause, *A*, and observe in the simulation whether the outcome, *B*, seemed likely or inevitable. This type of test could be used to explore the possible effects of an intervention, especially in forward causal reasoning (Woodward, 2006, and this volume). However, this type of simulation would appear to be of little value in cases where one were reasoning about the putative cause of a particular outcome that had already occurred, since the simulation of *A* and *B* would merely recapitulate the factual case that one was attempting to explain. Indeed, to the extent that reasoners regard such tests as evidence for '*A* caused *B*' rather than as an expression of the belief that '*A* caused *B*', they run the risk of being overconfident in the veracity of such beliefs (e.g. see Tetlock & Henik, 2005).

The second possibility goes a significant way towards getting around the 'problem of obviousness', whereby the simulation merely recapitulates the facts. In the second possibility, mental simulation might take the form of simulating *A* and observing whether *B* seems likely or inevitable, but crucially whilst negating other elements of the actual situation. By mentally altering factors other than the putative cause, its sufficiency across a set of close possible worlds could be ascertained. Such simulations might be important in situations where one knows that both *A* and *B* have happened, but one is unsure of the relation between the two events. By mentally varying (or intervening on) other factors in the scenario, one may be able to mentally probe the causal relationship between the two focal events, as well as the robustness or sufficiency of that relationship.

The third possibility, which represents the contraposition of the first possibility, would be to negate the outcome, *B*, and observe whether *A* would have to be negated for the simulation to be plausible. If so, one might increase one's confidence in the belief that *A* was sufficient for *B*, or at least that *A* significantly raised the probability of *B*. This type of test could be applied in retrospective assessments of causality since it does not merely reiterate the factual case. However, given that the antecedent would be conditional on the negated outcome, it might be difficult for people to employ this form of simulation. That is, conditional dependence in this case would be inconsistent with temporal order, which has been identified as an important cue to causality (Einhorn & Hogarth, 1986).

A Summary of Possibilities

The counterfactual possibilities brought to mind through mental simulations of causality for past episodes in which *A* and *B* occurred may be summarized as shown in

	B	$\neg B$
A	1 Confirmation by Factual Case	2 Sufficiency Violation by Counterfactual Case
$\neg A$	3 Necessity Violation by Counterfactual Case	4 Confirmation by Counterfactual Case

Figure 7.1 Factual and counterfactual possibilities in relation to the hypothesis '*A* caused *B*'.

Figure 7.1. That is, if one hypothesized that *A* caused *B*, then the factual co-occurrence of those events (i.e. cell 1) would be consistent with the hypothesis, as would counterfactual simulations in which the negation of *A* would result in the negation of *B* (i.e. cell 4). In contrast, counterfactual simulations in which either *A* occurs but *B* does not (i.e. cell 2) or *A* does not occur but *B* still occurs (cell 3) would diminish support for the same hypothesis. Specifically, simulations of cell 2 ought to diminish support for the hypothesis that *A* was sufficient to bring about *B*, and simulations of cell 3 diminish support for the hypothesis that *A* was necessary to bring about *B*. To put this in terms more conducive to the heuristics and biases framework, one might say that the sufficiency of *A* to bring about *B* may be assessed on the basis of how easy it is to imagine *A* occurring without *B*. The easier it is, the less likely the hypothesis would be to garner support from the simulation. Similarly, one might say that the necessity of *A* for bringing about *B* may be assessed on the basis of how easy it is to imagine the negation of *A* 'undoing' *B*'s occurrence. The easier it is, in this case, the more likely the hypothesis would be to garner support from the simulation.

To illustrate these ideas, consider the types of counterfactual arguments that were generated in the wake of the September 11, 2001, terrorist attacks on the United States. One claim that received much attention was that U.S. intelligence failures (*A* in our example) played a key role in allowing (and some might even go so far as to say *causing*) the 9/11 attacks to happen (*B* in our example). If by intelligence failures, we mean a set of events which we know did happen, such as not fully piecing together all of the available information that might have indicated that Al Qaeda was planning to attack the U.S. with hijacked airplanes, then the assertion of such a claim is itself merely a statement of the events in cell 1—intelligence failures occurred and, obviously, so did the 9/11 attacks. Of course, the problem of defining failure is more complex in real life and subject to less agreement that we might think (Lefebvre, 2004; Mandel, 2005b), but for illustrative purposes let's assume we agree that there were intelligence failures.

Having agreed on those facts, we may still disagree on the plausibility of the three types of counterfactuals represented by cells 2–4. For instance, some might point out that, as important as the intelligence failures might have been, they surely did not *necessitate* the

attacks. That is, had there been the same failures but no (or sufficiently weak) terrorist intent, 9/11 would not have occurred (an instance of cell 2). Proponents of that view would be inclined to argue that the intelligence failures by themselves were, therefore, insufficient to cause the attacks. Others might emphasize that even if the intelligence failures were minimized, the attacks would still have happened because it is nearly impossible to prevent terrorist attacks if the planners have sufficient resolve, which they apparently had (an instance of cell 3). Proponents of that view would be inclined to argue that such failures were not necessary causes of the attacks, even if they might have inadvertently served as enablers. Finally, advocates of the initial hypothesis that intelligence failures were a cause of the attacks might argue that, if only the intelligence had been 'better' (precisely how much better, and better in what ways?), the attacks would not have happened—or even that such attacks would not have been possible (an instance of cell 4). Thus, it appears that, while different observers might agree on the facts specified in cell 1, they may nevertheless disagree on the causal claim that explicitly refers to cell 1—namely, that intelligence failures caused the terrorist attacks. The plausibility of counterfactuals in cells 2–4 would seem to form part of the core argumentation for or against the putative cause.

Empirical Studies

Despite the widespread appeal of the notion that mental simulations of counterfactual scenarios play an important role in token-cause explanations, there have been relatively few studies that have directly addressed the issue. One of the more influential papers—by Wells & Gavanski (1989)—reported two scenario experiments in which the mutability of the negative outcome of the relevant episode was manipulated. In the low-mutability condition (in both studies), the negation of a focal antecedent event would not have changed the outcome, whereas in the high-mutability condition the same negation to the antecedent would have done so. For example, in one vignette a woman dies after eating an entrée that her boss ordered for her, which contained wine, an ingredient to which she was highly allergic. In the low-mutability version, the other dish the boss considered ordering for her also contained wine; whereas, in the high-mutability version, the other dish did not contain wine. Thus, had the boss chosen differently in the former version, it would not have made a difference, whereas it would have saved the woman's life in the latter version. The studies revealed that a significantly greater proportion of participants listed the target antecedent as both a cause and as a candidate for counterfactual mutation in the high-mutability condition than in the low-mutability condition. Participants that mutated the target antecedent were also more likely to list the same factor as a cause.

Wells & Gavanski concluded that their findings provided support for the idea that people arrive at causal explanations by using mental simulations of counterfactual scenarios. While the findings do support this interpretation, the studies constitute a fairly weak test. The stylized vignettes that they used left little else for participants to focus on as potential causes. For instance, in the vignette used in the first study just

described, there is little else for participants to focus on as potential causes other than the boss's ordering decision. This may be why nearly half of participants selected the boss's decision as the cause even in the low-mutability condition in which the other option would have made no difference, and perhaps also why the authors did not report the frequencies of other causal or counterfactual listings.

Indeed, subsequent research by N'gbala & Branscombe (1995) has shown that if vignettes are constructed with a broader range of explicit antecedents, participants focus on different factors in their causal and counterfactual responses. Specifically, participants tended to focus on necessary conditions for the occurrence of a particular outcome in their counterfactual explanations and sufficient conditions for that outcome's occurrence in their causal explanations. More recently, Mandel (2003b) found that, whereas counterfactual explanations tended to focus on antecedents that would have been sufficient to prevent a given type of outcome from occurring (e.g. a protagonist's death), causal explanations tended to focus on antecedents that played a role in how the actual outcome came about, especially if the antecedent was sufficient for the outcome as it actually occurred. For instance, in one study, participants read about a figure in organized crime, who unbeknownst to him, was poisoned with a slow-acting lethal dose that was sufficient to kill him. Prior to the poison taking effect, however, another assassin managed to kill the protagonist by ramming his car off the road. Thus, participants are presented with a case of causal overdetermination. The poison was sufficient to kill him and so was the car crash. After reading the scenario, participants were asked to list up to four causes of the protagonist's death and up to four ways his death might have been undone. Each of these listings was also rated in terms of its importance. Whereas participants regarded the car crash as the primary cause of the protagonist's death, they were most likely to counterfactually undo his death by mutating his involvement in organized crime. In fact, the importance assigned to a given antecedent in counterfactual and causal explanations were only weakly correlated. Thus, whereas their causal explanations focused on the factor that was sufficient to bring about the outcome as it actually occurred, their counterfactual explanations tended to focus on events that would have been sufficient to undo not only the actual outcome but also other inevitable outcomes that were categorically indistinct (i.e. all ways in which he was bound to be killed in the scenario).

Other studies (e.g. Davis et al. 1995; Mandel & Lehman, 1996; McEleney & Byrne, 2006) have shown that counterfactual and causal explanations also diverge in terms of the extent to which they are constrained by the perceived controllability of events. Counterfactual explanations of how an outcome might have been different tend to focus on antecedents that are controllable from a focal actor's perspective, whereas explanations of the cause of the same outcomes tend to focus on antecedents that would be predictive of similar outcomes in other episodes. For instance, Mandel & Lehman, (1996, Experiment 1) showed that, when participants were asked to explain how a car accident might have been undone from the perspective of the legally innocent victim, they tended to focus on controllable behaviors of that individual

(e.g. his choice of an unusual route home that day). In contrast, participants who were asked to generate causal explanations from the same victim's perspective were most likely to focus on the fact that the other driver was negligent (namely, he was under the influence of alcohol and ran a red light). Mandel & Lehman, (1996, Experiment 3) also found that whereas manipulations of antecedent mutability influenced participants' counterfactual responses, they had no effect on their causal responses. Other studies have also shown that counterfactual explanations are prone to focusing on controllable human behavior (Girotto, Legrenzi, & Rizzo, 1991; Morris, Moore, & Sim, 1999), and that controllable events are more likely than uncontrollable events to prompt the generation of spontaneous counterfactuals (McEleney & Byrne, 2006).

Only recently have some researchers attempted to directly manipulate the type of thinking that participants engage in prior to offering a causal judgment. Mandel (2003a) tested the hypothesis that counterfactual thinking about what *could have been* would have a stronger effect on participants' attributions than factual thinking about what actually *was*. For example, participants in Experiment 2 recalled an interpersonal conflict and were then instructed either to think counterfactually about something they (or someone else) might have done that would have altered the outcome or to think factually about something they (or someone else) actually did that contributed to how the outcome actually occurred. Participants rated their level of agreement with causality, preventability, controllability, and blame attributions, each of which implicated the actor specified in the thinking. Compared to participants in a baseline condition who did not receive a thinking directive, participants in the factual and counterfactual conditions had more extreme attributions regarding the relevant actor (either themselves or another individual with whom they had interacted) on the composite measure. Mean agreement, however, did not significantly differ between the factual and counterfactual conditions.

In a study using a similar thinking manipulation procedure, Mandel & Dhami (2005) found that sentenced prisoners assigned more blame to themselves and reported feeling guiltier for events leading up to their incarceration when they were first asked to think counterfactually rather than factually about those events. These effects are consistent with the view that counterfactual thinking prompts a focus on controllable actions (e.g. Girotto et al. 1991; Mandel & Lehman, 1996) and that, in hindsight, those mental simulations of what might have been often get translated into prescriptive judgments of what ought to have been (Miller & Turnbull, 1990). Mandel & Dhami (2005), however, did not examine the effect of thinking style on prisoners' causal explanations.

Summary

An overview of psychological research examining the correspondence of causal and counterfactual explanations does not offer compelling support for a very close coupling of the two in terms of explicit content. Some early studies claimed to have found support for the hypothesis that causal explanations are guided by counterfactual availability,

but as this research area developed and methodological limitations were overcome, support for this view has continued to wane. The findings, moreover, have tended to support an emerging view in which counterfactual and causal explanations are constrained by different selection criteria, which may serve distinct functional goals. For instance, counterfactuals that undo past negative outcomes seem to focus heavily on factors that were controllable from a focal actor's perspective (e.g. Davis et al. 1995; Mandel & Lehman, 1996) and, in cases of causal overdetermination, on factors that were able to undo not only the outcome as it actually occurred but an entire ad hoc category of outcome of which the actual serves as the prototype (Mandel, 2003b).

As I discuss in the next section, these findings support a functional account in which counterfactual explanations tend to elucidate ways that would have been sufficient to prevent unwanted outcomes or categories of outcome within a past episode. In contrast, the same studies have found that causal explanations tend to focus on factors that were perceived by participants to be sufficient under the circumstances for bringing about the outcome (or effect) as it actually was manifested in a past episode.

Although the studies reviewed, when taken together, do not support the view that the *explicit contents* of causal explanation are mainly filled in by the *explicit contents* of mentally available counterfactuals, it is important to note that they also do not rule out the possibility that causal explanations and causal thinking nevertheless depend in some fundamental way on counterfactual thinking. Stated differently, although the counterfactuals that people tend to explicate when asked to think about how a negative outcome might have been undone do not seem to correspond very closely with their causal explanations about the same outcomes, we cannot infer from those studies whether or not causal explanations might nevertheless rely upon some form of implicit counterfactual reasoning, as interventionist accounts, such as Woodward's (this volume), posit. Indeed, the foregoing analysis, which used Figure 7.1 as an organizing principle, suggests that implicit counterfactuals do play an important role in support and refutation of causal claims. Moreover, whereas Woodward (this volume) is reluctant to make any strong claim about the importance of counterfactual interventions in token-cause explanations, the preceding analysis clearly suggests that even token-cause explanations rely on counterfactual assessments. Thus, it remains an intriguing question why explicit counterfactuals do not better correspond to causal statements about the same episodes. Or, stated differently, it is intriguing why implicit and explicit counterfactuals do not correspond better. I will come back to this question towards the end of the chapter, after first outlining judgment dissociation theory.

4 Judgment Dissociation Theory: A Functional Account of Explanatory Divergence

According to most psychological accounts of counterfactual thinking, the 'if only' counterfactuals that people generate in mental simulations provide a heuristic test of

whether a factual antecedent was a necessary condition for a relevant outcome. But, as noted earlier, a necessity detector makes for a poor causal selection process, since there are simply too many necessary conditions for any given event, most of which would be deemed irrelevant (or worse) as the basis for a felicitous causal explanation. In the mid-1990s, I started developing an alternative, 'prevention focus' account (e.g. see Mandel & Lehman (1996)) that sought to better capture important differences between everyday causal explanations and everyday counterfactual ones. Subsequently, that account was further refined into 'judgment dissociation theory' (hereafter JDT; see Mandel, 2003b, 2005a), which developed the analysis of the functional dissociations between counterfactual and causal explanations.

JDT differs from the social psychological view dominant at the time it was proposed (what I'll call the 'standard view') in several important respects. First, JDT posits that people tend to focus on a sufficiency criterion when generating counterfactual explanations of the 'if only' or 'even if' variety, whereas the standard view assumed that people's counterfactuals tended to represent necessary conditions or causes. Second, JDT posits that everyday counterfactual statements, such as 'If only I had left home a few minutes earlier, I wouldn't have been stuck in this traffic jam', tend to focus on explaining how an unwanted outcome could (or at least might) have been prevented, whereas the standard view emphasizes the role that such thoughts play in explaining how the actual outcome was caused. Third, JDT posits that people tend to regard the antecedent of an 'if only' or 'even if' counterfactual conditional as an affirmation of a counterfactual state of the world (the affirmation of $\neg A$), whereas the standard view posits that people tend to regard the counterfactual conditional's antecedent as a denial (or at least mutation) of a factual state (i.e. the negation or mutation of A).

To illustrate these differences, it may be instructive to contrast the cells in Figure 7.1 with those in Figure 7.2, which alters the focal hypothesis from 'A caused B' to '$\neg A$ could have prevented B,' where the symbol '$\neg$' represents a mutated version of the relevant event. Note that, while cells 1 and 4 in Figure 7.2 still confirm the hypothesis, it is cell 4 rather than cell 1 that now is explicitly referenced by the prevention-oriented hypothesis. A comparison of the figures also reveals that, whereas cell 2 violates the

	B	$\neg B$
A	1 Confirmation by Factual Case	2 Necessity Violation by Counterfactual Case
$\neg A$	3 Sufficiency Violation by Counterfactual Case	4 Confirmation by Counterfactual Case

Figure 7.2 Factual and counterfactual possibilities in relation to the hypothesis '¬A could have prevented B'.

sufficiency of the putative cause in Figure 7.1, it violates the necessity of the putative preventer in Figure 7.2. Conversely, whereas cell 3 violates the necessity of the putative cause in Figure 7.1, it violates the sufficiency of the putative preventer in Figure 7.2. Of course, JDT does not propose that Figure 7.1 is wrong or that counterfactual thinking does not play an important role in causal explanation—it does not. It does, however, seek to highlight the equivocal interpretation of counterfactuals that have at times been treated unequivocally in the psychological literature (e.g. Roese & Olson, 1995; Wells & Gavanski, 1989). Moreover, JDT proposes that such counterfactual statements, more often than not, are aligned with a concern regarding the sufficiency of putative preventers of unwanted outcomes rather than the necessity of putative causes or enabling conditions of such outcomes.

Empirical studies

The prevention focus account and, subsequently, JDT have received support from several lines of research. For example, Mandel & Lehman (1996, Study 1) asked three groups of participants to either provide explanations of how a target outcome of an episode described in a vignette was caused, how it could have been prevented, or they were asked to generate if-only counterfactuals about the episode without reference in the query to either causality or preventability. The authors found that the counterfactual statements participants constructed were more closely aligned in content with the preventability explanations of other participants than with the causal explanations of other participants. In particular, both the counterfactual and preventability groups were more likely to provide responses that focused on a change to behavior that was controllable from a focal actor's perspective (e.g. the route the protagonist decided to take to drive home from work that day). By comparison, causal explanations were more likely to focus on the antecedent in the episode that would have been most predictive of the outcome in foresight and that, therefore, might also be useful in predicting similar types of outcomes in the future. Thus, while token-counterfactual explanations tended to focus on control, token-cause explanations tended to focus on predictability.

This content dissociation indicates an important functional dissociation between counterfactual and causal explanations. Clearly, both prediction and control are vital for an organism's success, but their relative emphasis in causal and counterfactual explanations appears to differ systematically. A similar dissociation was noted in Davis et al. (1995), who found that people often generated counterfactuals that undid a traumatic life event by changing something they could have controlled, even though few thought of the same actions as 'the cause' of the traumatic event. Taken together, such findings indicate that counterfactual explanations are closely aligned with manipulability notions of causation, which place an emphasis on human or agentic action (e.g. Collingwood, 1940).

Strictly speaking, such candidates for causal explanation seldom are sufficient causes because there are often several ways in which the causal impact of an event can be (and,

more importantly from a psychological perspective, could easily be perceived to be) thwarted intentionally or by fortuitous occurrence. The antecedents that people select, however, tend to be ones that, given the causal background in the particular episode, appear to be sufficient under the local circumstances of the episode in question (Mackie, 1974). Likewise, the controllable, counterfactual antecedents that people focus on in their counterfactuals and preventability explanations are ones that are easily seen as having been sufficient under the circumstances to prevent the outcome, even if they fail to meet a logical test of sufficiency. The *perceived* certainty of their effect, in hindsight, is indeed an important reason why counterfactuals are often so psychologically compelling.

A number of studies have shown that people define causality primarily in terms of sufficiency or a probabilistic variant of sufficiency in which the probability of the effect, given the presence of the putative cause, is relatively high. For example, Mandel & Lehman (1998, Experiment 1) asked participants to provide definitions of the words *cause* and *preventer*. They found that a majority of participants defined *cause* (71%) and *preventer* (76%) in terms of sufficiency (e.g. 'if the cause is present, the effect will occur'). By contrast, only a minority defined these concepts in terms of necessity (22% and 10% for *cause* and *preventer*, respectively; e.g. 'if the cause is absent, the effect won't occur'). Moreover, in a re-analysis of those data, Mandel (2005a) reported that all participants who had offered a necessity definition of a given term also provided a sufficiency definition of the same term. In other words, most participants were exclusively sufficiency focused and the remaining minority adopted a biconditional interpretation. None, however, was exclusively necessity focused in their interpretations of these terms. Mandel (2003b, Experiment 2) used a forced-choice task to investigate people's concept of causality. Participants were asked whether they thought the expression '*X* causes *Y*' means 'When *X* happens, *Y* also will happen' (i.e. *X* is sufficient to cause *Y*) or 'When *X* doesn't happen, *Y* also won't happen' (i.e. *X* is necessary to cause *Y*). Replicating the earlier findings based on an open-ended data-collection approach, a strong majority (81%) of participants selected the sufficiency interpretation.

Converging support comes from Goldvarg & Johnson-Laird (2001), who examined the types of possibilities that people view as being consistent with expressions like '*X* will cause *Y*'. For instance, in Experiment 1, they found that causal expressions were associated with the three possibilities implied by the notion of a sufficient cause for roughly half of the sample (i.e. *A* and *B*, $\neg A$ and *B*, and $\neg A$ and $\neg B$). The remainder indicated that the causal expressions were associated with the two possibilities implied by the notion of a necessary and sufficient cause (i.e. *A* and *B*, and $\neg A$ and $\neg B$). Thus, as in Mandel's (2005a) re-analysis, participants provided sufficiency or biconditional interpretations of causality. In addition, many studies of causal induction (namely, studies of how people infer causal laws, or type-causation, from statistical evidence) have shown that people assign greater weight to the sufficiency-relevant cause-present cases than to the necessity-relevant cause-absent cases when they judge the causal impact of a type of antecedent on a type of consequent (e.g. Anderson & Sheu, 1995;

Kao & Wasserman, 1993; Mandel & Lehman, 1998; Mandel & Vartanian, 2009; McGill, 1998; Schustack & Sternberg, 1981). Thus, several distinct lines of evidence converge in their support of JDT's proposal that causal and counterfactual explanations focus somewhat more on considerations of sufficiency than on considerations of necessity.

Although Mandel & Lehman (1996) emphasized the importance of control in counterfactual explanations (what they referred to as the controllability principle) and predictability in causal explanations (what they referred to as the covariational principle), later studies by Mandel (2003b) added an important caveat to the latter assertion. According to JDT, we must be clear about the manner in which causal explanations are sufficiency focused. That is, good candidates for token-cause explanations must not only be sufficient to bring about a particular type of event, they must also be sufficient under the circumstances to bring about the event *in fact*. As noted earlier, Mandel (2003b, Experiment 1) used a causal overdetermination scenario in which the protagonist was first lethally poisoned, but then was intentionally killed in a car crash, before the poison was able to yield its certain outcome. The poison was sufficient to kill the protagonist, but didn't. Even though it was temporally prior to the car crash episode, the latter was seen as a better causal explanation for the protagonist's death, presumably because it explained how he *actually* died. The poison only explains how the protagonist inevitably *would have* died if other events had not intervened to kill him first.

I designed these overdetermination scenarios mainly to test two hypotheses. First, I suspected that causal explanations could not be reduced to simple calculations of the conditional probability of the effect. Spellman (1997; also see Spellman, Kincannon, & Stose, 2005) had proposed an elegant model of token-cause explanation in which the antecedent from a set of putative causes that leads to the greatest increase in the probability of the effect, controlling for earlier antecedents, would be selected as the cause. If so, then participants should pick the poison (or the actor who administered it) as the cause, and not the car crash (or the driver who initiated it), since the former leads to a huge increase in the subjective probability of the protagonist's death, while the latter must lead to a negligible increase, given the virtual certainty of death by poisoning. As noted earlier, however, participants regarded the car crash episode as a superior causal explanation, even though they agreed that the poisoning led to the largest increase in the probability of the protagonist's death. Thus, the findings show that causal explanation cannot be reduced to 'explained variance' in the outcome. There are bounds on how much causal explanation simply tracks predictability or a probabilistic notion of sufficiency (cf. Mandel & Lehman, 1996).

The second aim of those experiments was to test the hypothesis that the proposed functional dissociation in causal and counterfactual explanation would manifest itself in terms of a systematic divergence in how 'the outcome' of an episode was to be defined. As already noted, JDT posits that causal explanations focus on a narrow view of the outcome; namely, on the actual outcome and not on merely inevitable outcomes that would have been functionally equivalent had they occurred instead, such as murder by

poisoning instead of murder by car crash. In other words, the actual outcome is not easily substitutable for similar types of outcomes in token-cause explanations—what I termed 'the actuality principle'. However, precisely because JDT posits that counterfactual explanations seek to identify sufficient preventers of an outcome, the functional value of this exercise would seem to be severely limited if such explanations were insensitive to 'merely inevitable' outcomes that were functionally indistinct from the actual outcome. Thus, it would not be a very satisfactory counterfactual explanation that undid the protagonist's death by car crash but allowed his death by poisoning. For this reason, I hypothesized that counterfactual explanations, particularly in cases of causal overdetermination, would favor a broad view of the outcome, whereby it was defined in terms of an 'ad hoc category' (Barsalou, 1983, 1991) in which the actual outcome would serve as the prototype—what I referred to as 'the substitutability principle'. Ad hoc categories, unlike natural categories, are usually based on short-term functional goals. Once those goals are achieved or no longer relevant, the category is 'disbanded'. 'Things to buy at the supermarket today' would be an example. As the example illustrates, not all ad hoc categories involve substitutable exemplars. Eggs don't substitute well for grapes. But, in counterfactual explanations, ad hoc categories are defined in terms of outcomes that are functionally substitutable, such as death by poisoning and death by car crash.

For instance, in the murder scenario, I expected that the outcome would be defined broadly as 'the protagonist's death' rather than narrowly as 'the protagonist's death by car crash'. If so, one might expect that counterfactual explanations in cases of multiple overdetermination would have to trace back further in time to an antecedent whose mutation could not only undo the actual outcome, but would also have prevented similar, probable outcomes too. In support of this prediction, it was found that participants' modal response was to undo the protagonist's life of crime—namely, the factor that motivated both the attempted murder by poisoning and the actual murder by car crash. In a related manner, Spellman & Kincannon (2001) found that, in cases of simultaneous overdetermination (e.g. two shooters shooting a victim at precisely the same moment), most participants offer explanations of how the outcome could have been prevented by undoing both rather than just one of the sufficient causes. Clearly, it is not satisfactory to simply replace 'death by two shots' with 'death by one'. As Hitchcock (this volume) notes, Lewis (1973/1986) seemed to be aware of the same substitutability requirement for a good counterfactual explanation; thus, he wrote:

> we certainly do not want counterfactuals saying that if a certain event had not occurred, a barely different event would have taken its place. They sound false; and they would make trouble for a counterfactual analysis of causation not just here, but quite generally. (1973/1986: 211)

As I see it, it is not so much that they sound false, as that fail to achieve their goal of offering a satisfactory account of undoing. JDT clarifies that the manner in which the outcome of an episode is construed is, in turn, shaped by functional considerations.

Summary

JDT is essentially a functional theory of explanation. It proposes that counterfactual and causal explanations serve different purposes and will, therefore, have some attributes that also predictably differ. Let us start, however, with the commonalities: both are presumed to serve goals that are generally adaptive. Moreover, both causal and counterfactual explanations are geared towards accounts of perceived sufficiency; or, more accurately, perceived sufficiency under the circumstances. In the former case, the putative cause should be sufficient under the circumstances to explain the occurrence of the actual outcome. In the latter case, the putative undoing antecedent should be sufficient under the circumstances to undo an ad hoc category of outcome of which the actual serves as the prototype. JDT is novel in this regard. No other theory of causal and counterfactual explanation makes predictions regarding how categorization processes will differ in the two cases. Indeed, although there is some research on the effect of causal thinking on categorization (e.g. Rehder & Hastie, 2001), there has surprisingly been virtually no research on the role of categorization in causal (and counterfactual) thinking.

5 Final Remarks

Mental simulation can play a role in formulating counterfactual and causal explanations. Given that the goodness of an explanation seems closely related to the plausibility of the scenario it conjures up, it is surprising that the effect of mental ease in scenario construction as a heuristic basis for judging explanatory quality has not received research attention. Clearly, research on this topic could help to elucidate the cognitive processes through which mental representations—generated through construction and/or recall—influence people's explanations and judgments.

Mental simulations pertinent to the causal explanation of a past event are indeed likely to be counterfactual, representing the expected effect of an intervention. However, the fact that such representations capture expectancies about interventions suggests that they are themselves predicated on causal knowledge, which may or may not have been predicated on counterfactual thought experiments. After all, a great deal of causal knowledge is acquired through cultural transmission. Even where such knowledge is predicated on counterfactuals, we face the perennial 'chicken-or-egg-first?' dilemma, and it would seem that, here too, the problem is non-reductive (see Woodward, 2003).

Although the emphasis in this chapter has been on showing how causal and counterfactual explanations systematically diverge, I have also cautioned the reader that this should not be interpreted as a rejection of the view that counterfactual thinking is central to causal reasoning. Here, we must distinguish between (explicit) counterfactual explanations of how unwanted events might have been prevented and the broader (most often implicit) class of counterfactual thoughts that might be

employed in causal reasoning and implied by causal statements. Both types of counterfactuals are examples of 'counterfactual availability', but they would seem to serve different purposes and have different likelihoods of being made explicit. Notably, the types of counterfactuals that Woodward (this volume) refers to in outlining his interventionist theory may be available to a causal reasoner without ever being articulated. They would seem to form part of the implicit understanding of what it means to say that *A* is a cause of *B*. In this sense, JDT ought to be regarded as an attempt to clarify that those causality-sustaining counterfactuals are not necessarily the same ones that sustain counterfactual explanations of how a past negative outcome might have been avoided. Unfortunately, my original exposition of JDT did not make this distinction clear, and thus it has probably, at times, been misread as a denial of the importance of counterfactual thinking in causal reasoning, which is not its intent.

JDT's 'actuality principle'—namely, that token-cause explanations are geared toward explaining the outcome as it actually occurred and not as it merely would or might have occurred—would also seem to suggest a way of moving toward an integration of interventionism and mechanistic accounts of causal reasoning (e.g. Ahn & Kalish, 2000; Salmon, 1984; Wolff & Song, 2003), since it suggests that, while intervention-counterfactuals might play a role in causal reasoning, causal explanations are guided by a concern over elucidating the mechanism that in fact brought about the outcome in the relevant episode. Here, too, I agree with Woodward (this volume) that such a project is an important one to which both psychology and philosophy may contribute.

References

Ahn, W. & Kalish, C.W. (2000) 'The role of mechanism beliefs in causal reasoning'. In: *Explanation and Cognition*, F.C. Keil & R.A. Wilson (eds), pp. 199–225. Cambridge, MA: Cambridge University Press.

Anderson, J.R. & Sheu, C.-F. (1995) 'Causal inferences as perceptual judgments', *Memory & Cognition* 23: 510–24.

Barsalou, L.W. (1983) 'Ad hoc categories', *Memory & Cognition* 11: 211–27.

—— (1991) 'Deriving categories to achieve goals'. In: *The Psychology of Learning and Motivation*, Vol. 27, G.H. Bower (ed.), pp. 1–64. New York: Academic Press.

Björnsson, G. (2006) 'How effects depend on their causes, why causal transitivity fails, and why we care about causation', *Philosophical Studies* 133: 349–90.

Byrne, R.M.J. (2005) *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

Cheng, P.W. (1993) 'Separating causal laws from causal facts: Pressing the limits of statistical relevance'. In: *The Psychology of Learning and Motivation*, Vol. 30, D.L. Medin (ed.), pp. 215–64. New York: Academic Press.

Collingwood, R.G. (1940) *An Essay on Metaphysics*. Oxford: Clarendon Press.

Davis, C.G., Lehman, D.R., Wortman, C.B., Silver, R.C., & Thompson, S.C. (1995) 'The undoing of traumatic life events', *Personality and Social Psychology Bulletin* 21: 109–24.

Dawes, R.M. (1996) 'Counterfactual inferences as instances of statistical inferences'. In: *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*, P.E. Tetlock & A. Belkin (eds), pp. 301–8. Princeton, NJ: Princeton University Press.

Einhorn, H.J. & Hogarth, R.M. (1986) 'Judging probable cause', *Psychological Bulletin* 99: 3–19.

Gärdenfors, P. (1988) *Knowledge in Flux*. Cambridge, MA: MIT Press.

Girotto, V., Legrenzi, P., & Rizzo, A. (1991) 'Event controllability in counterfactual thinking', *Acta Psychologica* 78: 111–33.

Goldvarg, E. & Johnson-Laird, P.N. (2001) 'Naïve causality: A mental model theory of causal meaning and reasoning', *Cognitive Science* 25: 565–610.

Grice, H.P. (1975) 'Logic and conversation'. In: *Syntax and Semantics 3: Speech Acts*, P. Cole & J.L. Morgan (eds), pp. 41–58. New York: Academic Press.

Halpern, J. & Pearl, J. (2005) 'Causes and explanations. A structural model approach. Part II: Explanations', *British Journal for the Philosophy of Science* 56: 889–911.

Hart, H.L.A. & Honoré, A.M. (1985) *Causation in the Law* (2nd edition). Oxford: Clarendon.

Heider, F. (1958) *The Psychology of Interpersonal Relations*. New York: Wiley.

Hesslow, G. (1988) 'The problem of causal selection'. In: *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, D.J. Hilton (ed.), pp. 11–32. Brighton, UK: Harvester Press.

Hilton, D. J. (1990) 'Conversational processes and causal explanation', *Psychological Bulletin* 107: 65–81.

—— McClure, J.L., & Slugoski, B.R. (2005) 'The course of events: counterfactuals, causal sequences, and explanation'. In: *The Psychology of Counterfactual Thinking*, D.R. Mandel, D.J. Hilton, & P. Catellani (eds), pp. 44–60. New York: Routledge.

—— & Slugoski, B.R. (1986) 'Knowledge-based causal attribution: The abnormal conditions focus model', *Psychological Review* 93: 75–88.

Johnson-Laird, P.N. & Byrne, R.M.J. (2002) 'Conditionals: A theory of meaning, pragmatics, and inference', *Psychological Review* 109: 646–78.

Kahneman, D. (1995) 'Varieties of counterfactual thinking'. In: *What Might Have Been: The Social Psychology of Counterfactual Thinking*, N.J. Roese & J.M. Olson (eds), pp. 375–96. Mahwah, NJ: Erlbaum.

—— & Miller, D.T. (1986) 'Norm theory: Comparing reality to its alternatives', *Psychological Review* 93: 136–53.

Kahneman, D. & Tversky, A. (1982a) 'The simulation heuristic'. In: *Judgment under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, & A. Tversky (eds). pp. 201–8. New York: Cambridge University Press.

—— —— (1982b) 'Variants of uncertainty', *Cognition* 11: 143–57.

—— Slovic, P., & Tversky, A. (1982) (eds) *Judgment under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

—— & Varey, C.A. (1990) 'Propensities and counterfactuals: The loser that almost won', *Journal of Personality and Social Psychology* 59: 1101–10.

Kao, S.-F. & Wasserman, E.A. (1993) 'Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19: 1363–86.

Keil, F.C. (2006) 'Explanation and understanding', *Annual Review of Psychology* 57: 227–54.

Killeen, P.R. (2001) 'The four causes of behavior', *Current Directions in Psychological Science* 10: 136–40.

Lefebvre, S. (2004) 'A look at intelligence analysis', *International Journal of Intelligence and Counterintelligence* 17: 231–64.

Lewis, D. (1973/1986) 'Causation', *Journal of Philosophy* 70: 556–67. Reprinted with postscripts in his *Philosophical Papers,* Vol. 2, Oxford: Oxford University Press, 1986, pp. 159–213. (Page references are to the reprinted version.)

Lombard, L.M. (1990) 'Causes, enablers, and the counterfactual analysis', *Philosophical Studies* 59: 195–211.

McEleney, A. & Byrne, R.M.J. (2006) 'Spontaneous counterfactual thoughts and causal explanations', *Thinking & Reasoning* 12: 235–55.

McGill, A.L. (1998) 'Relative use of necessity and sufficiency information in causal judgments about natural categories', *Journal of Personality and Social Psychology* 75: 70–81.

Mackie, J.L. (1974) *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.

Mandel, D.R. (2003a) 'Effect of counterfactual and factual thinking on causal judgments', *Thinking & Reasoning* 9: 246–65.

—— (2003b) 'Judgment dissociation theory: An analysis of differences in causal, counterfactual, and covariational reasoning', *Journal of Experimental Psychology: General* 132: 419–34.

—— (2005a) 'Counterfactual and causal explanation: From early theoretical views to new frontiers'. In: *The Psychology of Counterfactual Thinking*, D.R. Mandel, D.J. Hilton, & P. Catellani (eds), pp. 11–27. New York: Routledge.

—— (2005b) 'Threats to democracy: A judgment and decision making perspective', *Analyses of Social Issues and Public Policy* 5: 209–22.

—— (2007) 'Differential focus in causal and counterfactual thinking: Different possibilities or different functions?', *Behavioral and Brain Sciences* 30: 460–1.

—— (2008) 'Violations of coherence in subjective probability: A representational and assessment processes account', *Cognition* 106: 130–56.

—— & Dhami, M.K. (2005) '"What I did" versus "What I might have done": Effect of factual and counterfactual thinking on blame, guilt, and shame in prisoners', *Journal of Experimental Social Psychology* 41: 637–45.

—— & Lehman, D.R. (1996) 'Counterfactual thinking and ascriptions of cause and preventability', *Journal of Personality and Social Psychology* 71: 450–63.

—— —— (1998) 'Integration of contingency information in judgments of cause, covariation, and probability', *Journal of Experimental Psychology: General* 127: 269–85.

—— & Vartanian, O. (2009) 'Weighting of contingency information in causal judgment: Evidence of hypothesis dependence and use of a positive-test strategy', *Quarterly Journal of Experimental Psychology* 62: 2388–408.

Miller, D.T. & Turnbull, W. (1990) 'The counterfactual fallacy: Confusing what might have been with what ought to have been', *Social Justice Research* 4: 1–19.

Morris, M.W., Moore, P.C., & Sim, D.L.H. (1999) 'Choosing remedies after accidents: Counterfactual thoughts and the focus on fixing "human error"', *Psychonomic Bulletin & Review* 6: 579–85.

N'gbala, A. & Branscombe, N.R. (1995) 'Mental simulation and causal attribution: When simulating an event does not affect fault assignment', *Journal of Experimental Social Psychology* 31: 139–62.

Rehder, B. & Hastie, R. (2001) 'Causal knowledge and categories: The effects of causal beliefs on categorization, induction, and similarity', *Journal of Experimental Psychology: General* 130: 323–60.

Roese, N.J. & Olson, J.M. (1995) 'Counterfactual thinking: A critical overview'. In: *What Might Have Been: The Social Psychology of Counterfactual Thinking*, N.J. Roese & J.M. Olson (eds), pp. 1–55. Mahwah, NJ: Erlbaum.

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

Schustack, M.W. & Sternberg, R.J. (1981) 'Evaluation of evidence in causal inference', *Journal of Experimental Psychology: General* 110: 101–20.

Spellman, B.A. (1997) 'Crediting causality', *Journal of Experimental Psychology: General* 126: 323–48.

—— & Kincannon, A. (2001) 'The relation between counterfactual ("but for") and causal reasoning: Experimental findings and implications for jurors' decisions', *Law and Contemporary Problems* 64: 241–64.

—— Kincannon, A.P., & Stose S.J. (2005) 'The relation between counterfactual and causal reasoning'. In: *The Psychology of Counterfactual Thinking*, D.R. Mandel, D.J. Hilton, & P. Catellani (eds), pp. 28–43. New York: Routledge.

Tetlock, P.E. & Belkin, A. (1996) 'Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives'. In: *Counterfactual Thought Experiments in World Politics: Logical, Methodological, and Psychological Perspectives*, P.E. Tetlock & A. Belkin (eds), pp. 3–38. Princeton, NJ: Princeton University Press.

—— Henik, E. (2005) 'Theory- versus imagination-driven thinking about historical counterfactuals: Are we prisoners of our preconceptions?'. In: *The Psychology of Counterfactual Thinking*, D.R. Mandel, D.J. Hilton, & P. Catellani (eds), pp. 199–216. New York: Routledge.

—— Visser, P.S., Singh, R., Polifroni, M., Scott, A., Elson, S.B., Mazzocco, P., & Rescober, P. (2007) 'People as intuitive prosecutors: the impact of social-control goals on attributions of responsibility', *Journal of Experimental Social Psychology* 43: 195–209.

Trabasso, T. & Bartolone, J. (2003) 'Story understanding and counterfactual reasoning', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 904–23.

Tversky, A. & Kahneman, D. (1973) 'Availability: A heuristic for judging frequency and probability', *Cognitive Psychology* 5: 207–32.

Wells, G.L. & Gavanski, I. (1989) 'Mental simulation of causality', *Journal of Personality and Social Psychology* 56: 161–69.

Wilson D. & Sperber D. (2004) 'Relevance theory'. In:*Handbook of Pragmatics*, G.Ward & L. Horn (eds), pp.607–32. Oxford: Blackwell.

Wolff, P. & Song, G. (2003) 'Models of causation and the semantics of causal verbs', *Cognitive Psychology* 47: 276–332.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

—— (2006) 'Sensitive and insensitive causation', *The Philosophical Review* 115: 1–50.

8

Counterfactual Availability and Causal Judgment

Christopher Hitchcock

In this paper I survey the psychological literature on *counterfactual availability*. While humans have the capacity to engage in counterfactual reasoning, we do not do so indiscriminately. Rather, we tend to entertain specific kinds of counterfactual hypotheses. Psychologists have studied the sorts of factors that influence the counterfactuals we entertain. I argue that we can expect these factors to exert a related influence on causal judgments, and connect this thesis with a number of discussions from the philosophical literature on causation.

1 Counterfactuals and Causation

David Lewis (1973/1986) and others have tried to analyze causation in terms of counterfactuals. The simplest such theory would equate causation with counterfactual dependence: *C* is a cause of *E* just in case if *C* hadn't occurred, *E* would not have occurred.[1] There are a number of problems with this simple theory. One problem that has been widely discussed in the philosophical literature is the problem of pre-emption. In cases of pre-emption, one event *C* causes another event *E*, but *C* also pre-empts event *D* which would, in the absence of *C*, have caused E. In such cases, *C* is a cause of *E*, even though *E* would still have occurred if *C* had not. For example, in the movie *Gosford Park*, the victim William McCordle is first poisoned and then posthumously stabbed by a second would-be murderer. The poison killed him, even though he would have died (from the stabbing) if he had not been poisoned. A number of attempts have been made to modify the simple counterfactual theory to accommodate cases of pre-emption,[2] but we will not concern ourselves with the details here.

[1] Two qualifications are needed for even this simple theory to be remotely plausible: *C* and *E* must be events that both occur, and they must be suitably independent of one another. If, for example, the occurrence of *E* logically entails the occurrence of *C*, then there may be counterfactual dependence without causation.

[2] See Hitchcock (2001, 2007) for my own attempts.

A different sort of worry is that a counterfactual analysis of causation may get the conceptual priority in the wrong order. It may be that we need to understand counterfactuals in terms of causation (see e.g. Kvart, 1986, Edgington, this volume). Alternatively, it may be that causation and counterfactuals are related as siblings, both descended from the same parent concept (Maudlin, 2004).

In psychology, a number of researchers (such as Kahneman & Tversky, 1982) have suggested that we use counterfactual reasoning when making causal judgments. To judge whether *C* was a cause of *E*, we imagine that *C* had not occurred, and run a simulation to determine if *E* would have been absent as well. Mandel (this volume) presents some reasons for doubting this psychological model of causal reasoning.

These worries notwithstanding, it seems clear that there is some kind of close relationship between causation and counterfactuals. Counterfactual dependence provides a pretty good *prima facie* test for causation. It is used, for example, as a criterion for causation in British and American common law. In this paper, I will not presuppose any particular counterfactual analysis of causation, or any particular model of the psychology of causal reasoning exploiting counterfactuals. I will argue, however, that the relationship between causation and counterfactuals is sufficiently close that the sorts of factors that influence counterfactual reasoning will also have a parallel effect on our causal judgments.

2 Two Tribes Look at Counterfactuals

Both philosophers and psychologists have shown a great deal of interest in counterfactuals, although the focus of their attention has been markedly different. Philosophers have been primarily interested in questions about the logic and semantics of counterfactuals. The logic of counterfactuals concerns inferences that can be made using counterfactuals. Philosophers (e.g. Lewis, 1973) have observed, for example, that the counterfactual conditional does not always permit strengthening of the antecedent. That is, from the counterfactual 'if *A* had occurred, *C* would have occurred', we cannot necessarily infer 'if *A* and *B* had occurred, *C* would have occurred'. This is in contrast to the indicative conditional, where strengthening of the antecedent is logically valid. The semantics of counterfactuals is concerned with what makes counterfactuals true. This can be approached formally, often using a 'possible worlds' framework. To evaluate the counterfactual 'if *A* had occurred, *C* would have occurred', when spoken or written in the actual world, we move to the 'closest' possible world to the actual world in which *A* occurs, and see if *C* holds there. It is of course a nice question what makes some possible worlds closer than others. According to Lewis (1973), the closest possible *A*-world is the one with the greatest overall similarity. It turns out, however, that the criteria of similarity needed to get the counterfactuals to come out right are far from intuitive (Lewis, 1979). Dorothy Edgington discusses some of these difficulties in her contribution to this volume.

Psychologists have typically held a different set of interests in counterfactuals. One set of issues concerns how we reason counterfactually. What are the cognitive mechanisms that we employ, what more general cognitive capacities are involved, and how are these implemented in the brain? A second set of issues involves the development of counterfactual thinking in children, a topic that is explored in the contributions of Beck, Riggs, & Burns and of Perner & Rafetseder, in this volume. Third, psychologists have explored the question of what triggers counterfactual thinking. The general consensus is that counterfactual thinking is triggered by negative affect, and serves the purpose of providing guidance for improvement in the future (see e.g. Roese, 1997). Fourth, they have been interested in the emotional consequences of counterfactual thinking, typically feelings of regret. Indeed, in a number of experiments, regret is used as a marker for counterfactual thinking. Finally, psychologists have been interested in the question of which counterfactual possibilities we actually entertain. There are a great many counterfactuals that we *could* think about and evaluate. If someone asked me what would happen if my desk were made of Jello, I could correctly reason that my computer would sink to the floor. There is nothing inherently unthinkable about this counterfactual. But as a matter of fact, I would never entertain such a counterfactual, unless someone specifically directed me to do so (or if I wanted to give an example of a counterexample that no one would ever bother to entertain). Using the terminology of Tversky & Kahneman (1973), some counterfactual possibilities are more readily *available* than others. It is this last aspect of the psychology of counterfactual thinking that I will focus on here.

Prima facie, the psychological question of which counterfactual hypotheses we actually entertain looks orthogonal to the sorts of questions that have exercised philosophers. One is reminded here of the linguists' distinction between competence and performance. Linguists, especially syntacticians, have been primarily concerned with linguistic competence: which sentences would an informant recognize as grammatically correct sentences of her native language? Performance will not match competence. Real speakers often speak in sentence fragments, and make performance errors due to a variety of disturbing factors. Moreover, any given speaker will utter only a tiny fraction of the syntactically available sentences of her language during her lifetime. Similarly, we might expect the set of counterfactuals that I actually entertain to be a small subset of the counterfactuals that I could recognize as true.

I want to suggest that there may be a stronger interaction between the psychology and the semantics of counterfactuals, and that this interaction will affect concepts like causation that are closely connected with counterfactuals. Here are some possible mechanisms by which such interactions might occur. First, if the antecedent of a counterfactual may be satisfied in different ways, counterfactual availability may affect how we imagine it to be satisfied. Recall the basic idea employed in most semantics for counterfactuals: the truth of the counterfactual 'If A had occurred, then B would have occurred' is determined by what happens in the 'closest' possible world in which A occurs. If A may be realized in different ways in different possible worlds, it may be that

the 'closeness' of these worlds corresponds to the availability of the alternative ways of realizing *A*. This mechanism is particularly relevant for the sorts of counterfactuals that are germane to causation. Recall that those counterfactuals have the form 'if *C* had not occurred, then *E* would not have occurred', where *C* and *E* are specific events. Typically, there will be many alternatives to *C*, or ways for *C* not to occur. Counterfactual availability may affect which alternative to *C* we take to occur in the nearest possible world where *C* fails to occur. This may, in turn, affect whether *E* follows from the supposition that *C* does not occur.

Second, if a claim that does not explicitly have the form of a counterfactual is to be understood in counterfactual terms—either precisely, in the form of a philosophical analysis, or more loosely, as suggested by Kahneman & Tversky (1982)—then the counterfactual paraphrase may have a tacit 'subject to availability' clause. For example, it may be that the proper paraphrase of '*C* caused *E*' is not 'if *C* had not occurred, *E* would not have occurred', but rather 'there is an available alternative to *C*, such that if that alternative had occurred, *E* would not have occurred'. If, as a matter of psychology, we actually use a counterfactual test for causation, then our causal judgments will almost certainly behave this way. If there is no available alternative to *C* in which *E* fails, then we will not judge that *E* would not have occurred if *C* had not occurred, and hence we will not judge that *C* caused *E*.

Alternately, if causation is logically prior to counterfactuals, as Edgington (this volume) argues, and if causal judgments are also psychologically prior, we might find the explanation running in the other direction. That is, it may be that the factors that contribute to counterfactual availability do so only indirectly. Perhaps these factors contribute directly to our causal judgments, and we are prone to entertain counterfactual hypotheses in which those events that we judge to be causes, as well as those features of events that we judge to be causally efficacious, are mutated. Then our causal judgments will also guide us in judging that the effect would not have occurred in the hypothetical scenario.

In practice, these mechanisms will tend to produce the same correlation between our causal judgments and our counterfactual reflections: we will judge that *C* causes *E* when there is a psychologically available alternative to *C* in which *E* does not occur. My strategy in the sections that follow will be to canvass a variety of factors that psychologists have identified as influencing counterfactual availability. It is natural to suspect that these factors will also affect causal judgments. I will then point to a number of discussions in the philosophical literature that suggest that these factors do influence causal judgments in just the ways that we would expect.

3 Factors that Influence Counterfactual Availability

Psychologists have identified a number of factors that can influence the availability of counterfactual alternatives. We will survey the four most prominent.

Focal and Background Factors

Factors that are the focus of attention are more readily changed than factors that are part of the presupposed background. Read (1985) presents two experiments that illustrate this effect. In one, subjects are presented with a vignette about two people (creatively named 'A' and 'B') playing a card game (essentially a simplified version of poker). Subjects were then asked to complete statements beginning 'A would have won if . . . ', or 'A would have lost if . . . '. Because of the logic of the game, the counterfactual supposition 'A would have won' is equivalent to 'B would have lost', and vice versa. But by framing the question from A's perspective, A became the focus, and B part of the background. Subjects predominantly completed the sentences by modifying A's hand, even though the same result could have been achieved by modifying B's hand. In another experiment, subjects were presented a vignette in which a driver, Helen, gets herself into a dangerous situation facing an oncoming car. The vignette focused on Helen, rather than on the driver of the other car. Subjects were then asked how an accident might be avoided. A majority of subjects responded by proposing hypothetical actions that Helen could undertake, rather than actions the other driver could perform, even though the description of the situation made it clear that it would be difficult for Helen to avoid the accident.

Proximate and Remote Possibilities

Unrealized possibilities that are proximate in space and time to events that actually occurred seem to be more readily available than counterfactual possibilities that are more remote in space and time. When thinking about what might have been, we have a tendency to focus on near misses. In one classic experiment (Kahneman & Tversky, 1982), subjects were given a vignette in which two people, Mr. Crane and Mr. Tees, are caught in traffic on the way to the airport, and arrive 30 minutes after the scheduled departure of their flights. However, while Mr. Crane's flight left on time, 30 minutes before his arrival, Mr. Tees' flight was delayed 25 minutes, so that he missed his flight by only 5 minutes. Subjects were then asked who they thought would be more upset. The overwhelming majority said that Mr. Tees would be more upset. Note that in this experiment, the negative affect associated with feelings of regret is being used as a marker for counterfactual thinking. Recently, however, Kühberger, Perner, & Ziesel (MS) have attained results that shed some doubt on the conclusions drawn from this experiment. They used the same vignette, but in contrast to Kahneman & Tversky, they used a between-subjects design. Subjects were given either the story about Mr. Crane, or the story about Mr. Tees, and were asked to rate how upset he would be. In this experiment, subjects gave the same ratings, regardless of which vignette they had seen. Kühberger et al. attribute the difference in result to different methods employed by the subjects to assess the level of regret experienced by the two travelers. They argue that the within-subjects design employed by Kahneman & Tversky invites subjects to apply a (possibly false) folk-psychological theory, while the between-subjects design of Kühberger et al. invites subjects to simulate the regret that they would feel in the

situation described. However, other studies employing a between-subjects design have also suggested an effect of proximity on counterfactual availability. Miller & McFarland (1986) presented subjects with a vignette in which there is a plane crash. A man survives the crash, but then dies while trying to get help. In one version, he dies 75 miles from the nearest town, in the other, a quarter mile away. Subjects were then asked how much in damages his heirs should receive from the airline company. Somewhat surprisingly, subjects awarded higher damages in the case where the man nearly made it to safety. This is explained by the fact that the availability of the alternative in which he survives increased the negative affect associated with his death.

Medvec, Madey, & Gilovich (1995) found that Olympic bronze medalists showed greater satisfaction in their accomplishment than Olympic silver medalists. They attributed this counter-intuitive finding to the bronze medalists' reflecting on the possibility that they might have failed to earn a medal altogether, while the silver medalists reflected on the possibility that they might have won the gold. Since this study is based on the actual responses of the athletes, there can be no issue of whether regret is assessed by a folk theory or by simulation.

If the purpose of counterfactual thinking is to consider ways in which we might improve our performance in the future, then it makes sense to focus on near misses, since this is where the prospects for improvement are greatest.

Exception and Routine

Routine events are more readily available as counterfactual alternatives than exceptional events. When constructing alternatives to actuality, we more naturally replace exceptional events with their routine alternatives than the converse. Kahneman & Tversky (1982) presented subjects with a vignette in which Mr. Jones was killed in a traffic accident on the way home from work when he was hit by a drug-crazed truck-driving teenager who ran a red light. In one version of the story, Mr. Jones left home from work early in order to run some errands; in another version, he deviated from his normal route home in order to enjoy the scenic drive along the lake. Subjects were informed of both possibilities; for example, subjects who were given the first version of the story were also told that Mr. Jones occasionally took a scenic route home, but that he took his normal route home that day. Subjects were then told that Jones' family and friends often found themselves thinking 'if only ...'; the subjects were then told to complete the 'if only ...' statements. Those subjects who were given the first version of the story were strongly inclined to say 'if only Jones had left at his regular time', rather than 'if only Jones had taken the scenic route'; whereas subjects who were given the second version of the story were inclined to say 'if only Jones had taken his regular route home' rather than 'if only Jones had left early to run errands'. Wells, Taylor, & Turtle (1987) attained similar results.

Trabasso & Bartolone (2003) recently proposed an alternative explanation of these findings. In the vignettes used in the experiments, the exceptional factors received more explanation than the routine factors. This is understandable; we typically do not

think it necessary to explain why Jones departed at his usual time, while it is necessary to explain why he took an unusual route home. Trabasso & Bartolone argue that this difference in explanation accounts for the difference in counterfactual availability. This explanation appears to be closely related to the influence of focus on availability.

Ideals and Violations

Ideal values of attributes or variables tend to be highly available; hence people are strongly inclined to construct counterfactual alternatives by replacing non-ideal elements with ideal elements. In the experiment described earlier, Read (1985) presented subjects with a vignette in which two people are playing cards. One person wins the hand, and subjects are invited to complete the sentence, 'The outcome would have been different if . . . '. Subjects were much more strongly inclined to strengthen the losing hand than to weaken the winning hand. In another experiment, subjects were told about two tennis players, Tom and Jim, who both were eliminated from a tournament when they lost a point on a tie-breaker. Tom lost when his opponent served an ace, while Jim lost when he made an unforced error. Subjects judged that Jim was more likely to spend time thinking about the possibility that he might have won the match.

4 Causal Judgment

We will now consider how these various factors that can influence counterfactual availability may also influence causal judgments. In general, if causal judgments are closely tied to counterfactual reasoning, we would expect those factors that are most easily mutated to create counterfactual possibilities to be most readily judged as causes. We will explore several applications of this idea.

Focal and Background Factors

A number of authors have commented on the role of focal stress in causal claims. Dretske (1977) provides the following example. Consider the following two causal claims:

> Susan's *stealing* the bicycle caused her to be arrested.
> Susan's stealing *the bicycle* caused her to be arrested.

We might judge that the first of these claims is true, while the second is false. Intuitively, it is *that* she stole, and not *what* she stole, that caused her to be arrested. (We assume that the police do not have a particular penchant for arresting bicycle thieves as opposed to other kinds of thieves.)

Dretske's example raises a prima facie metaphysical puzzle. It would seem that the event picked out by the nominal 'Susan's stealing the bicycle' is the same, regardless of where the stress is placed. So if causation is a relation between events, it is hard to see how one of these claims could be true while the other is false. Dretske argued that this

kind of case shows that causation is not a relationship between events, but between more fine-grained relata that he called 'event allomorphs'. Yablo (1992) suggests something similar. He proposes that the nominals do indeed pick out different events. The events differ with respect to their essential and accidental properties. Susan's *stealing* the bicycle is an event that is essentially a stealing, and only accidentally involves a bicycle, while Susan's stealing *the bicycle* is an event that essentially involves a bicycle, and is only accidentally a stealing. Lewis (1973/1986, 1986) makes remarks that suggest that he would also endorse this line of response. If this is right, then it may be that our tendency to entertain counterfactuals in which features of a situation that are in focus are changed results from our tendency to mutate the essential features of the events that are described. For example, when the stress is on the word *bicycle*, the bicycle is brought into the foreground, and the nominal refers to an event that essentially involves a bicycle. Then it may be that our tendency to mutate this feature of the situation in counterfactual thinking is a direct result of the event's essence, and only indirectly a result of focus.

I think, however, that a more natural and ontologically less profligate account results from taking the counterfactuals to be prior.[3] The different placement of focal stress in the two claims makes different counterfactual alternatives available. In the first claim, the focus is placed upon the stealing (the action), while Susan (the agent), and the bicycle (the thing acted upon), remain in the background. The factors that are most readily mutated to produce counterfactual alternatives are those that are in the foreground. Thus the counterfactual alternatives that are most readily available are those in which the actor, Susan, and the thing acted upon, the bicycle, remain unchanged, but the action is changed. These will be alternatives in which Susan performs some other action with the bicycle, such as purchasing it, renting it, or borrowing it. In these alternative scenarios, Susan is not likely to be arrested; hence the counterfactual test for causation is met. By contrast, when the stress is placed upon 'the bicycle', it (the thing acted upon) becomes the focus, while Susan (the agent) and the stealing (the action) remain in the background. In this case, the counterfactual alternatives that are most readily available are those in which Susan steals something other than the bicycle. In these alternative scenarios, Susan is just as likely to be arrested, and the counterfactual test for causation is not met.

Proximate and Remote Factors

David Lewis appears to deny that proximity should affect our evaluation of counterfactuals:

> wouldn't the minimal change to get rid of an event be one that replaces it with a barely different event? Not so; a similarity theory needn't suppose that just any sort of similarity we can think of has non-zero weight ... we certainly do not want counterfactuals saying that if a

[3] See Hitchcock (1996) for a much more detailed discussion.

> certain event had not occurred, a barely different event would have taken its place. They sound false; and they would make trouble for a counterfactual analysis of causation not just here, but quite generally. (Lewis, 1973/1986: 211)

In order to see the problem, consider the canonical causal claim that Socrates's drinking hemlock caused his death. To assess whether this claim is true, we consider what would have happened if the event of Socrates drinking hemlock hadn't occurred. The worry is that if we replace that event by one in which Socrates drinks hemlock at a nearby place or nearby time, Socrates's death will still occur; hence we would not judge that his drinking hemlock caused his death. And it is easy to see that the problem is quite general: almost nothing would count as a cause if we allowed proximity to play this kind of role in the evaluation of counterfactuals. But there is a problem here. A possible situation in which Socrates drinks hemlock at a slightly different time or place is not clearly one in which the event of his drinking hemlock does not occur. So in changing Socrates's actual drinking of hemlock to a proximate drinking of hemlock, we have not clearly applied the counterfactual test.

If the relevant spatial or temporal location is stated explicitly, then it does seem appropriate to allow proximity to play a role. Suppose, for example, it was claimed that Socrates's drinking hemlock at 6:00 p.m. caused his death. Then it seems reasonable to consider what would have happened if he had instead consumed the hemlock at 5:59 or 6:01. We might imagine that for some strange reason, Socrates would not have died if he had consumed the hemlock at noon, but this is not enough to make the causal claim true. In ascribing Socrates's death to his consumption of hemlock at 6:00 p.m., we seem to be implying that the exact time of consumption was important, and that death would not have occurred if he had consumed the hemlock even slightly earlier or slightly later. Note that focus is also playing a role here. When the time of drinking is explicitly stated, it becomes a subject of focus, and hence alternatives in which the time of drinking is different become readily available. By contrast, if the time of drinking is not stated explicitly in the causal claim, then we do not tend to generate counterfactual alternatives by mutating the time of drinking.

Exception and Routine

When asked to identify the causes of some event, we tend to identify events that were unusual or unexpected. We find, for example, in Hart & Honoré (1985) the proposal that causes are, by definition, departures from the normal course of events that lead to unusual outcomes that would not have occurred otherwise. This idea seems to be at the root of the commonly made distinction between causes and background conditions. If there is a forest fire, we tend to regard the carelessly tossed cigarette as a cause, while the presence of oxygen, although equally necessary for the fire, is relegated to the status of a background condition. Since the tossing of the cigarette involves a departure from the routine course of events, we readily imagine the case where it is absent, and hence we easily judge it to be a cause of the fire. By contrast, the presence of oxygen is

routine, and the alternative in which no oxygen is present is less readily available to us, so we are less willing to judge that the oxygen is a cause of the fire.[4] If the context changes so as to make the presence of oxygen exceptional, for example, if the fire takes place in a room in a high tech lab where oxygen is normally evacuated, then we would be more willing to judge the presence of oxygen to be a cause of the fire.

Ideals and Violations

Since we tend to replace departures from the ideal with the ideal alternative rather than vice versa when constructing counterfactual scenarios, we would expect that departures from the ideal will be more readily judged as causes. This would help to explain divergent judgments in certain cases of causation by omission.[5] To use a standard example, suppose that nobody watered the flowers and they died. Now consider the following two claims:

> The gardener's failure to water the flowers caused them to die.
> The Queen's failure to water the flowers caused them to die.

We are more inclined to accept the former claim than the latter, even though the flowers would not have died if *either* the gardener or the Queen had watered them. The gardener's failure to water the flowers involves the violation of a norm—the gardener was *supposed* to water the flowers—so the alternative in which he waters the flowers is readily available. By contrast, the Queen's failure to water the flowers did not involve any violation of a norm, so the alternative in which she waters the flowers is not readily available.

5 More on Norms and Violations

Joshua Knobe and I have recently done some empirical work on the effect of norms on causal judgments.[6] Here is an experiment reported in an earlier paper by Knobe & Fraser (2008). Subjects were presented with the following vignette:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take pens, but faculty members are supposed to buy their own.
>
> The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist repeatedly e-mailed them reminders that only administrators are allowed to take the pens.
>
> On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later, that day, the receptionist needs to take an important message ... but she has a problem. There are no pens left on her desk.

[4] For more detailed discussion, see Hitchcock (2007), section 9.
[5] See, for example, the discussions in Beebee (2004) and McGrath (2005).
[6] Described more fully in Hitchcock & Knobe (2009, MS)

Subjects were then asked to rate the extent to which they agreed or disagreed with one of the following statements using a 7-point scale (1 = strongly disagree, 7 = strongly agree):

> Professor Smith caused the problem.
> The administrative assistant caused the problem.

Subjects who were given the first statement gave an average rating of 6.2, indicating agreement, while subjects asked the second question gave an average rating of 2.8, indicating disagreement ($p < .001$). This suggests that the norm forbidding faculty members from taking the pens led subjects to view Professor Smith's action as a cause of the problem, but not the administrative assistant's action, just as predicted.

Joshua Knobe and I wanted to test some potential alternative explanations for why norms might influence our causal attributions in this way. One explanation is that somehow the presence of the norm is distorting the subject's response, and hence the response does not accurately reflect the subjects' competence with the concept of causation. There are different ways such an explanation might go. Perhaps subjects are conflating the questions of causation and blame. Or perhaps there is a pragmatic explanation: while it is strictly true that the administrative assistant caused the problem, it would be infelicitous to say so, since the context suggests that the causal inquiry is being made for the purpose of assigning blame (Driver, 2008, suggests such a response). Or perhaps subjects are engaging in rationalization. They have an immediate negative affective reaction to Professor Smith's action, which biases them in favor of causal assessments that will underwrite attributions of blame to him (Alicke, 1992, proposes a model along these lines).

We can test these potential explanations by seeing whether the effect of norms on causal judgments is present when blame is no longer appropriate, for example if the outcome is good, or if moral assessment does not apply because no agents are involved.

In order to do this, we presented subjects with two new scenarios. Here is the first:

> An intern is taking care of a patient in a hospital. The intern notices that the patient is having some kidney problems. Recently, the intern read a series of studies about a new drug that can alleviate problems like this one, and he decides to administer the drug in this case.
>
> Before the intern can administer the drug, he needs to get the signature of the pharmacist (to confirm that the hospital has enough in stock) and the signature of the attending doctor (to confirm that the drug is appropriate for this patient). So he sends off requests to both the pharmacist and the attending doctor.
>
> The pharmacist receives the request, checks to see that they have enough in stock, and immediately signs off.
>
> The attending doctor receives the request at the same time and immediately realizes that there are strong reasons to refuse. Although some studies show that the drug can help people with kidney problems, there are also a number of studies showing that the drug can have very dangerous side effects. For this reason, the hospital has a policy forbidding the use of this drug for kidney problems. Despite this policy, the doctor decides to sign off.

Since both signatures were received, the patient is administered the drug. As it happens, the patient immediately recovers, and the drug has no adverse effects.

Subjects were then asked to rate their agreement or disagreement with one of the following two statements on the same seven-point scale:

> The pharmacist's decision caused the patient's recovery.
> The attending doctor's decision caused the patient's recovery.

The survey was presented online, with subjects being recruited through advertisements on several websites. Over 3,000 subjects participated in the survey. Subjects who were asked the first question gave an average rating of 2.53, while those asked the second question gave an average rating of 3.94 ($p < .001$). Subjects were thus inclined to judge that the pharmacist's decision was not a cause of the patient's recovery, but were (on average) somewhat inclined to judge that the doctor's decision was a cause (recall that 4 was the mid-point on the scale). Since the outcome was good, there cannot be any question of *blaming* the attending doctor for the outcome. This suggests that it is in fact because the attending doctor violated a norm that subjects were more strongly inclined to judge that his decision was a cause.

In the same survey, we also tested subjects on the following scenario:

A machine is set up in such a way that it will short circuit if both the black wire and the red wire touch the battery at the same time. The machine will not short circuit if just one of these wires touches the battery. The black wire is designated as the one that is supposed to touch the battery, while the red wire is supposed to remain in some other part of the machine.

One day, the black wire and the red wire both end up touching the battery at the same time. There is a short circuit.

Subjects were asked to rate their agreement or disagreement with one of the following claims:

> The fact that the red wire touched the battery caused the machine to short circuit.
> The fact that the black wire touched the battery caused the machine to short circuit.

Subjects who were asked about the red wire gave an average rating of 4.89 on the seven-point scale, while those asked about the black wire gave an average rating of 2.67 ($p < .001$). Here the norm in question is not a moral norm, but rather a norm of proper functioning. There can be no question of literally blaming the red wire, since an inanimate object cannot be the subject of moral evaluation. Nor does it seem plausible that the vignette evoked a strong affective response of the sort that might trigger a desire to attribute responsibility to the red wire. Moreover, the vignette does not provide any additional explanatory information about why the red wire touched the battery, so the effect cannot be accounted for in terms of availability created through explanation in the model of Trabasso & Bartolone (2003). Once again, this suggests that it is because the location of the red wire violated a norm that subjects judged it to be a cause of the short circuit.

Incidentally, our subject pool contained over 300 subjects who identified themselves as professors of philosophy, or holding a PhD in philosophy. This sub-group gave essentially the same responses as the other subjects (although all of their ratings tended to be a little higher). In particular, they showed statistically significant differences in their ratings between cases where a norm was violated and cases where no norm was violated. Presumably, trained philosophers are capable of distinguishing questions of causation from questions of blame, so these results cast further doubt on the hypothesis that the difference in ratings stem from subjects' conflation of causation and blame.

6 Conclusion

The philosophical discussions surveyed in Section 4 strongly suggest that the sorts of factors that influence counterfactual availability also affect causal judgments. Specifically, we are more strongly inclined to judge that *C* is a cause of *E* when alternatives to *C* in which *E* does not occur are counterfactually available. This could be explained naturally if subjects employ a counterfactual test for causation. Alternatively, it may be that these factors directly influence our causal judgments, which in turn influence the counterfactual hypotheses we actually entertain, and which also guide our evaluation of these counterfactuals.

Further empirical work would need to be done to determine whether ordinary subjects make causal judgments which accord with the intuitions marshaled in these philosophical discussions. Empirical studies on the effect of norms on causal judgments suggest that in the case of ideals and violation, at least, they do. Moreover, this research strongly suggests that it is norm violations *per se* that are influencing subjects' causal judgments, and not a conflation of causation and blame. I have argued that a strong psychological connection between counterfactual availability and causal judgments provides the best explanation of these results.[7]

References

Alicke, M. (1992) 'Culpable Causation', *Journal of Personality and Social Psychology* 63: 368–78.

Beebee, H. (2004) 'Causing and Nothingness', in Collins, Hall, & Paul (2004), pp. 291–308.

Collins, J., Hall, N., & Paul, L. (2004) (eds) *Causation and Counterfactuals*. Cambridge MA: MIT Press.

Dretske, F. (1977) 'Referring to Events', *Midwest Studies in Philosophy* 2: 90–9.

[7] I would like to thank Peter Godfrey-Smith, Alison Gopnik, Dennis Hilton, Josh Knobe, David Lagnado, David Mandel, Jim Woodward, audience members at the McDonnell workshop on Causal and Moral Cognition (California Institute of Technology), the University of Southern California, the workshop on the Origins and Functions of Causal Thinking IV (Venice, Italy), the Workshop on Causal and Counterfactual Understanding (University of Warwick), and the Workshop on Counterfactuals (Erasmus University, Rotterdam).

Driver, J. (2008) 'Attributions of Causation and Moral Responsibility', in Sinnott-Armstrong (2008), pp. 423–39.

Hart, H.L.A. & Honoré, T. (1985) *Causation in the Law* (2nd Edition). Oxford: Oxford University Press.

Hitchcock, C. (1996) '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.

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

—— & Knobe, J. (2009) 'Cause and Norm', *Journal of Philosophy* 106: 587–612.

—— —— (MS) 'The Influence of Norms and Intentions on Causal Attributions'. Unpublished Manuscript, California Institute of Technology.

Kahneman, D. & Miller, D. (1986) 'Norm Theory: Comparing Reality to its Alternatives', *Psychological Review* 80: 136–53.

—— & Tversky, A. (1982) 'The Simulation Heuristic'. In: *Judgment Under Incertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, & A. Tversky (eds), pp. 201–10. Cambridge: Cambridge University Press.

Knobe, J. & Fraser, B. (2008) 'Causal Judgment and Moral Judgment: Two Experiments', in Sinnott-Armstrong (2008), pp. 441–7.

Kühberger, A., Perner, J., & Ziesel, D. (MS) 'Lay Theories of Counterfactual Thinking'. Unpublished Manuscript, University of Salzburg.

Kvart, I. (1986) *A Theory of Counterfactuals*. Indianapolis: Hackett.

Lewis, D. (1973) *Counterfactuals*. Cambridge, MA.: Harvard University Press.

—— (1973/1986) 'Causation', *Journal of Philosophy* 70: 556–67. Reprinted with postscripts in Lewis (1986a), pp. 159–213. (Page references are to the reprinted version.)

—— (1979) 'Counterfactual Dependence and Time's Arrow', *Noûs* 13: 455–76. Reprinted in Lewis (1986a), pp. 32–52.

—— (1986) 'Events', in Lewis (1986a), pp. 241–69.

—— (1986a) *Philosophical Papers, Volume II*. Oxford: Oxford University Press.

McGrath, S. (2005) 'Causation by Omission', *Philosophical Studies* 123: 125–48.

Maudlin, T. (2004) 'Causation, Counterfactuals, and the Third Factor', in Collins, Hall, & Paul (2004), pp. 419–43.

Medvec, V., Madey, S., & Gilovich, T. (1995) 'When Less is More: Counterfactual Thinking and Satisfaction among Olympic Athletes', *Journal of Personality and Social Psychology* 69: 603–10.

Miller, D. & McFarland, C. (1986) 'Counterfactual Thinking and Victim Compensation: A Test of Norm Theory', *Personality and Social Psychology Bulletin* 12: 513–19.

Read, D. (1985) 'Deteminants of Relative Mutability'. Unpublished Manuscript, University of British Columbia.

Roese, N. (1997) 'Counterfactual Thinking', *Psychological Bulletin* 121: 133–48.

Sinnott-Armstrong, W. (2008) (ed.) *Moral Psychology, Volume 2: The Cognitive Science of Morality*. Cambridge, MA: MIT Press.

Trabasso, T. & Bartolone, J. (2003) 'Story Understanding and Counterfactual Reasoning', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 904–23.

Tversky, A. & Kahneman, D. (1973) 'Availability: A Heuristic for Judging Frequency and Probability', *Cognitive Psychology* 5: 207–32.

Wells, G., Taylor, B., & Turtle, J. (1987) 'The Undoing of Scenarios', *Journal of Personality and Social Psychology* 53: 421–30.

Yablo, S. (1992) 'Cause and Essence', *Synthese* 93: 403–49.

9

The Role of Counterfactual Dependence in Causal Judgements

Peter Menzies

1 Introduction

What is the connection between counterfactuals and actual or token causation? Both philosophers and psychologists have addressed this question, though with different foci of interest. Philosophers have tended to focus on giving the semantics of causal sentences by stating their truth conditions, whereas psychologists have tended to focus on the patterns of thinking and reasoning by which people arrive at their causal judgements. Despite the difference in their immediate focus, it should be possible for philosophers and psychologists to learn from each others' insights. For knowing how people arrive at their causal judgements should shed light on the meanings they give to these judgements, and conversely, knowledge of the semantics of causal sentences should help to illuminate the way they engage in causal reasoning and judgement.

The consensus view among philosophers is that there is some connection between causation and counterfactuals, but it is doubtful whether the meaning of causal judgements can be exhaustively explained in terms of counterfactuals. Much of this philosophical discussion has centred on the correctness of David Lewis's (1973b [1986]) famous counterfactual theory of causation.[1] Lewis analysed causation in terms of the notion of counterfactual dependence between events. Roughly speaking, *e* counterfactually depends on *c* just in case if *c* had not occurred, *e* would not have occurred. So, at the heart of Lewis's analysis is the famous 'but-for' test of causation. A great deal of philosophical ink has been spilled on trying to show that his theory cannot account for our causal judgements about pre-emption examples. (For example, Hall, 2004; Menzies, 1996; Schaffer, 2000 [2004]; Woodward, 1990). These are cases in which two potential causes of some effect are present: one of them succeeds in bringing about the effect and in the process pre-empts or cuts short the causal efficacy of the other.

[1] Strictly speaking, Lewis advanced another two counterfactual theories of causation, which developed his original theory in new ways. See his (1986, Postscript E) and (2000 [2004]). In this paper I shall discuss only his earliest theory.

The classic example involves two assassins, a main assassin and a back-up assassin, hired to kill a prominent individual: the main assassin carries out the execution, pre-empting the action of the back-up. However, the victim's death does not counterfactually depend on the main assassin's action, since if the main assassin had not acted, the victim would have died anyway through the back-up assassin's action. It seems that examples of this kind pose particularly recalcitrant difficulties for Lewis's theory.

Among psychologists a similar consensus has arisen that while causal reasoning and counterfactual reasoning are closely connected, they are distinct forms of reasoning. The psychological research has focused on determining the factors that prompt people's counterfactual judgements and those that prompt their causal judgements. It has been found that these are different sets of factors. (See Mandel, 2003). For example, in one experiment subjects are presented with a vignette about an individual who takes an unusual route home, but is hit by a drunk during the drive. When asked to think about relevant counterfactuals about this individual, most respond with the counterfactual, 'If he hadn't taken the unusual route home, he wouldn't have been involved in the accident'. At the same time, most people judge that the actions of the drunk driver were the principal cause of the accident. More generally, psychological research has concluded that counterfactual reasoning is useful in identifying the enabling conditions of some event, but not useful in identifying its causes. (See Byrne, 2005: ch. 5.)

My aim in this paper is to argue that philosophers and psychologists have been premature in dismissing the possibility that the causal concept is analytically tied to the concept of counterfactual dependence. I shall argue that if we understand the notion of counterfactual dependence in a suitably enriched way, we can see that some examples that purport to show the difference between causation and counterfactual dependence do not in fact show this. Philosophers have had difficulty seeing this because they have taken for granted a certain thesis about counterfactuals: namely, the thesis that a counterfactual with a true antecedent has trivial truth conditions so that it is true just in case its consequent is true. I shall argue that this standard assumption about counterfactuals is false, which means that counterfactuals with true antecedents must be given substantial truth conditions. In stating these truth conditions, I shall draw on work by David Lewis (1986, Postscript C) and James Woodward (2006) on insensitive causation. They claim that insensitive causation requires a counterfactual with a true antecedent to hold not only in the actual world, but also in nearby worlds that do not depart much from the actual world. To specify these worlds more precisely I shall draw on work in cognitive psychology on counterfactual availability, or the conditions under which people spontaneously generate counterfactuals or evaluate particular counterfactuals as true. I shall rely, in particular, on studies by Kahneman & Tversky (1982), Kahneman & Miller (1986), and Hart & Honoré (1985) that attend to the important role that the concepts of norms, normal conditions, and interventions play in our counterfactual reasoning. With these non-trivial truth conditions in place, I shall argue that a conception of counterfactual dependence that employs non-trivial

counterfactuals with true antecedents is a more apt tool for analysing the concept of causation.

In detail, here is my plan of action. In Section 2 I shall explain the basic features of David Lewis's counterfactual theory of causation. In Section 3 I shall outline some of the problems thought to afflict this theory, concentrating not on the problems of pre-emption that philosophers have been preoccupied with, but on issues more likely to resonate with psychologists, issues concerning the distinction between enabling conditions and causes, and the distinction between positive and negative causes. In Section 4 I shall motivate and justify the particular manner in which I propose to enrich the concept of counterfactual dependence by providing substantial truth conditions for counterfactuals with true antecedents. In Section 5 I shall explain how understanding causation in terms of the enriched notion of counterfactual dependence can explain our intuitive judgements about the problematic examples described in Section 3.

2 Lewis's Counterfactual Theory of Causation

Among philosophers, the best-known counterfactual theory of causation is David Lewis's (1973b [1986]) theory. (For other kinds of counterfactual theory, see Mackie, 1974; Woodward, 2003.) It will be useful to focus our discussion on Lewis's theory, as it is a mature, well-articulated theory. Lewis claims that causation can be analysed in terms of counterfactual dependence, which he defines in these terms:

> An event *e* *counterfactually depends* on event *c* if and only if (i) if *c* had occurred, *e* would have occurred; and (ii) if *c* had not occurred, *e* would not have occurred.

Counterfactuals are understood in his semantics in terms of similarity relations between possible worlds (Lewis, 1973a). In this semantics, each world is assigned a system of spheres of possible worlds that is centred on it; and these spheres carry information about the overall similarity between the world at the centre and other worlds. In the system centred on possible world *w*, the smaller the sphere the more similar to *w* must be the worlds that fall within it. Lewis imposes certain formal constraints on a system of spheres, the most important for our purposes being a centring principle that states that in the system of spheres centred on the world *w*, the innermost sphere contains only *w*. So if *w* is the actual world, this principle implies that there is no world as similar to the actual world as the actual is to itself. Given sets of spheres obeying these constraints, Lewis is able to state the truth conditions for counterfactuals. Let us say that within a system of spheres, the smallest sphere that allows '*P*' to be true is the smallest *P*-permitting sphere. Then the counterfactual 'If *P* were the case, then *Q* would be the case' is true in world *w* if and only if '$P \supset Q$' is true throughout the smallest *P*-permitting sphere.[2] Another way to express this is to say that the counterfactual is

[2] These are the conditions for non-vacuously true counterfactuals. The counterfactual is vacuously true if there is no *P*-permitting sphere.

true if and only 'Q' is true in all the worlds most similar to w in which 'P' is true, or 'Q' is true in the closest P-worlds.

A point that will be important for our later discussion is that the centring principle that Lewis adopts means that a counterfactual with a true antecedent has very simple truth conditions. If we are considering a system of spheres centred on the actual world, then if 'P' is true in the actual world, the counterfactual 'If P were the case, Q would be the case' is true if and only '$P \supset Q$' is true in the actual world, since the actual world is the smallest P-permitting sphere. Consequently, since we almost invariably apply the definition of counterfactual dependence to actual occurrent events, the first counterfactual (i) 'If c had occurred, e would have occurred' is automatically true. For in these circumstances 'If c had occurred, e would have occurred' has the truth conditions of the corresponding material conditional 'c occurred $\supset$ e occurred', and this material conditional is true when its antecedent and consequent are true. The upshot of this is that the truth conditions of the definition of counterfactual dependence can be simplified as follows:

An occurrent event e *counterfactually depends* on an occurrent event c if and only if if c had not occurred, e would not have occurred.

In applying this semantics to particular counterfactuals, Lewis notes that the concept of similarity is to be understood as a technical notion rather than an intuitive notion of the kind that might be expressed in off-hand judgements of similarity. If similarity were understood in the latter way, the theory would be open to serious objection. For example, Kit Fine (1975) pointed out that when similarity is interpreted intuitively, the counterfactual 'If Nixon had pressed the button, there would have been a nuclear holocaust' comes out false instead of true. For the worlds in which Nixon presses the button that are most intuitively similar to the actual world are ones where the effects of his action are neutralized so that the world continues as it actually did. Lewis (1979 [1986]) gives a precise description of the technical notion of similarity that is to govern the interpretation of counterfactuals relevant to analysing causation. Called the *standard, non-backtracking interpretation,* it prescribes that in the typical circumstances in which the temporal asymmetry of counterfactuals prevails, the closest worlds to the actual world for evaluating a counterfactual 'If P were the case, Q would be the case', where P is about states of affairs at time t, are as follows: they are worlds in which P is true; worlds which are exactly like the actual world at all times before a transition period beginning shortly before t; which conform to the actual laws after t; and which during the transition period differ no more from the actual world than they must to permit P to hold.[3] These worlds hold fixed the past up until shortly before the time at

[3] Lewis settles on a different analysis of similarity: see his (1979 [1986]). The analysis described here is an analysis that he dismisses because it builds in the temporal asymmetry of counterfactuals by fiat; and so is unsuitable in his view because it rules out the possibility of backwards causation. But since backwards causation is not an issue for us, we can safely settle on his second-best analysis.

which the antecedent is imagined to hold; and so rule out the kind of backtracking reasoning that tries to infer how the past had to have been different in order for the counterfactual antecedent to hold.

Lewis introduces two further conditions into his analysis of causation. First, he stipulates that counterfactual dependence amounts to causal dependence only when it holds between wholly distinct events. Sometimes a counterfactual dependence can hold between events that are not wholly distinct from each other. For example, it may be true that if a flag had not been red, it would not have been crimson; and if I had not torn the envelope, I would not have torn the return address written on its back. But these are not genuine causal dependences since the flag's being crimson is not a distinct state of affairs from its being red, and my tearing the envelope is not a distinct event from my tearing the return address written on its back. So Lewis gives this definition:

An event *e causally depends* on an event *c* if and only if (i) *c* and *e* are wholly distinct; and (ii) *e* would not have occurred if *c* had not occurred.

Secondly, Lewis believes that causation is a transitive relation, and since the relation of causal dependence, as just defined, is not transitive, he analyses causation in terms of the ancestral of causal dependence. His final analysis is as follows:

An event *c caused* an event *e* if and only if there is a chain of causal dependences from *c* to *e*.

Lewis deploys this reference to chains of counterfactual dependence in his treatment of pre-emption, but this reference will not be relevant to our discussion, which will eschew pre-emption examples. The significant point is that the direct causal dependence of *e* on *c* counts as a chain of causal dependence, a chain consisting with one link, which suffices on this analysis for the existence of causation between *c* and *e*.

3 Some Problems Facing Lewis's Theory

Lewis's theory has faced a great deal of criticism by philosophers, which has mostly revolved around whether it is able to handle cases of pre-emption. Setting these cases to one side, I want to concentrate on some counterexamples to the theory that connect up with issues discussed in the psychological literature about causation and counterfactuals concerning the distinction between enabling conditions and causes, and between positive and negative causes.

One set of common criticisms of Lewis's theory concerns the fact that his theory glosses over a distinction we draw between causes and enabling conditions. (In Menzies, 2004, this is called the problem of profligate causes.) Consider this familiar example:

Example 1: Birth and Death
A man is born and much later in his life dies in a car accident. The cause of his death was the car accident and his birth is merely an enabling condition of his death.

Lewis's theory fails to mark the distinction in causal status between the car accident and the man's birth since it counts them both as straightforward causes. For it is true that if the car accident had not occurred, the man would not have died; and also true that if the man had not been born, he would not have later died. Since all the events involved are distinct, these are causal dependences that imply the existence of causal relations.

Lewis's theory's over-generation of causes is exacerbated by the fact that it is supposed to cover causation by omissions and absences as well as actions and events. Lewis says that it is a virtue of his theory that it respects commonsense causal judgements by accommodating absences as causes and effects. After all, we say such things as parents' failure to vaccinate children caused their illness, the failure of a government to conserve water supplies caused a drought, and a doctor's omission to prescribe a drug caused a patient's death. However, the admission of absences as causes compounds the problem of the over-generation of causes. Consider the following example.

Example 2: The Absence of Meteor Strike

I am writing this paper at my computer. If, however, there were nerve gas in the air, or if I were attacked with flame-throwers, or struck by a meteor shower, I would not be writing this essay. But it is absurd to say that the absence of nerve gas in the air, the absence of flame-thrower attacks, and the absence of a meteor strike are causes of my writing this essay.

Another example familiar in the philosophical literature illustrates the difficulty faced by Lewis's theory in distinguishing between genuine and spurious causes among omissions.

Example 3: The Pot Plant

A pot plant, which is otherwise healthy, requires regular watering during hot weather. A gardener whose job it is to water the plant fails to do so during the hot spell and the plant dies. If the gardener had watered the plant, it would have survived. So his failure counts as a cause of the plant's death. But it also true that if the Queen of England had watered the plant, it would have survived. It is absurd, nonetheless, to say that the Queen's failure was a cause of the plant's death.

Again Lewis's theory cannot explain the discriminating causal judgements we make in these cases.

Lewis was aware of these difficulties but did not regard them as decisive. Like many philosophers, he tended to dismiss the distinction between causes and enabling conditions as unimportant. He defended his theory against such counterexamples by referring to Grice's pragmatic theory of conversation implicature. He maintained that it is literally true that any event or absence on which an effect counterfactually depends is a cause of the effect; but it is not always conversationally appropriate to cite it as a cause. He writes: 'There are ever so many reasons why it might be inappropriate to say something true. It might be irrelevant to the conversation, it might convey a false hint, it might be known already to all concerned' (2000 [2004]: 101). However, Lewis provided scant details of the way Grice's conversational maxims are meant to apply to particular examples. Which maxims are relevant? How are they to be applied?

Moreover, there is a question of whether Grice's principles are especially well suited to explaining the specific causal judgement in question. Grice's maxims are very general principles of rationality applied to information exchange. Yet the principles that lie behind our judgements about the examples above seem to be particular to causal judgements. As general principles of rational information exchange, Grice's maxims miss out these particular causation-specific principles.

A number of other problems in a similar vein to the ones above have recently emerged in the philosophical literature. Carolina Sartorio (2010) has described some problems she calls the Prince of Wales problems.

Example 4: The Problem of Unwanted Positive Causes
The Queen of England has asked the Prince of Wales to water her plants regularly in the afternoon. But the Prince's priorities are to eat oaten biscuits instead of watering the plants and, consequently, the plant dies. The Prince's failure to water the plants, not his eating the oaten biscuits, caused the plant's death.

Sartorio argues that part of the point of admitting absences and omissions as causes is that they are not replaceable by positive causes: we take the Prince's omission to be a cause of the plant's death because the positive action he performed instead of watering the plants would not have the same causes and effects. But it appears that Lewis's theory implies that the Prince's eating the oaten biscuits is as much a cause of the plant's death as his failure to water the plants. If we add to the example the supposition that the Prince had prioritized his activities so that he would have watered the plants if he had not eaten the oaten biscuits, it follows that his eating the oaten biscuits counts as a cause of the plant's death. For if he had not eaten the biscuits, he would have watered the plant and it would have survived. Since the Prince's eating the biscuits is distinct from the plant's death, we have a causal relation. But this is one too many causal relations when we take into account that a causal relation also exists between the Prince's failure to water the plant and its death. (For the plant would have survived if the Prince had watered the plants, and the Prince's failure to water the plant and its death are distinct.) This is what Sartorio calls the problem of unwanted positive causes. It is a problem for Lewis's theory precisely because it undermines one of the reasons Lewis gives for admitting absences as causes: to wit, that there are not adequate positive events to replace the absences in the explanation of their effects.

Sartorio notes that the theory also has the converse problem: it generates unwanted negative causes as well. Consider this example:

Example 5: The Problem of Unwanted Negative Causes
The Prince of Wales's eats oaten biscuits instead of watering the plant. But he eats so many oaten biscuits that he gets a stomach ache. The Prince's eating too many oaten biscuits, not his failure to water the plant, caused his stomach ache.

Sartorio points out that Lewis's theory mistakenly implies that the Prince's failure to water the plants is a cause of his stomach ache. For the Prince's failure to water the plant

is distinct from his stomach ache and there is a counterfactual dependence between them: if the Prince had watered the plant instead of eating the oaten biscuits, he would not have got the stomach ache. Again, we have one too many causes, but this time we have an unwanted negative cause. As Sartorio observes, this problem is more serious than the converse problem, which generates an unwanted positive cause only when it is assumed that if the positive cause had not occurred, the negative cause would not have occurred. In contrast, the problem of unwanted negative causes arises whenever there is an event whose occurrence is precluded by the occurrence of the positive cause. Consider all the activities precluded by the Prince's eating the oaten biscuits: instead of eating the oaten biscuits, he could have watered the Queen's plant, talked to the Duke, walked in his gardens, and so on. Then Lewis's theory implies that corresponding to each of these actions is an omission that counts as cause of the Prince's stomach ache, since for each action it is true that if the Prince had performed that action rather than omitting it, he would not have got a stomach ache. Hence, it follows that the Prince's failure to water the plant, his failure to talk to the Duke, his failure to walk in his garden, and so on all count as causes of his stomach ache. This is far too many causes.

4 Deviant and Default Counterfactuals

How is the problem of the over-generation of causes to be tackled? One obvious strategy is to augment Lewis's theory by adding a further condition to his analysis of causation in order to eliminate the unwanted causal candidates. In this section I propose to follow something like this strategy. But instead of adding further conditions to his analysis of causation in terms of counterfactual dependence, I propose to strengthen his definition of counterfactual dependence in a natural way. Lewis's definition of counterfactual dependence, applied to states of affairs *c* and *e* including positive occurrences like events as well as negative occurrences like absences and omissions, requires the truth of two counterfactuals:

(i) If *c* were to obtain, *e* would obtain;
(ii) If *c* were not to obtain, *e* would not obtain.

Lewis's imposes a centring principle on his semantics for counterfactuals from which it follows that the first counterfactual (i) is true when its antecedent and consequent are true (Lewis, 1973a: 14). This principle implies that in cases in which the states of affairs *c* and *e* obtain, the first counterfactual (i) is trivially true with the result that the definition of counterfactual dependence effectively reduces to the second counterfactual (ii). I believe that Lewis is mistaken in imposing this centring principle on his semantics. When this principle is suspended, substantive truth conditions can be given to counterfactuals with true antecedents; and this has the effect of making counterfactual dependence a much stronger condition in so far as it requires the holding of two non-trivial counterfactual conditions. As we shall see in the next section, the account of

causation in terms of this strengthened condition of counterfactual dependence goes a long way to explaining our intuitions about the various causal scenarios described in the last section.

The centring principle that Lewis imposes on his semantics for counterfactuals states that the system of spheres centred on the actual world (or any given world) contains a smallest, innermost sphere containing only the actual world (or the given world). Lewis justifies this principle on the grounds that no world is as similar to the actual world as the actual world is to itself. It follows from this principle that when a counterfactual has a true antecedent, the smallest antecedent-permitting sphere is the innermost sphere that contains only the actual world. This implies that the truth conditions of counterfactuals with true antecedents reduce to those of the corresponding material conditionals; and so a counterfactual with a true antecedent like (i) is true if and only the material conditional '*c* obtains $\supset$ *e* obtains' is true. From this it follows that inferences of the following form are valid: *c* obtains and *e* obtains; therefore, if *c* were to obtain, *e* would obtain. But this is a fallacious inference. It would be absurd to reason that the match would light if it were struck on the basis of the fact that the match is struck and the match lights. For these states of affairs might obtain as a pure coincidence, which would hardly justify the counterfactual claim, which implies a strong law-like connection between the antecedent and consequent states of affairs. The absurdity of the inference pattern is even more evident in cases in which one tries to reason from two completely unrelated states of affairs such as the state of affairs of the Prime Minister scratching his nose and the state of affairs of Big Ben striking the hour a second later to the corresponding counterfactual claim that if the Prime Minister were to scratch his nose, Big Ben would strike a second later.

Lewis's centring principle has other unacceptable consequences such as the fact that it validates the rule of inference Strengthening of the Antecedent for counterfactuals with true antecedents. That it validates the rule is evident from the fact that counterfactuals with true antecedents reduce to material conditionals and the rule is valid for material conditionals. But this result conflicts with naïve judgements about the validity of inferences of this form. The inference from 'If the match were struck, it would light' to 'If the match were struck in the absence of oxygen, it would light' is just as bad when the antecedent 'The match is struck' is true as when it is false.

These results give us reason to be suspicious of this semantic centring principle. To be sure, the principle sounds plausible enough when we judge similarity in an offhand, casual way. How could there be a world that is more similar to the actual world than the actual world itself? But examined more closely, the principle is not so plausible. The applicability of the principle requires unreasonably fine-grained judgements of similarity and difference between possible worlds. We have to be able to judge similarity between worlds down to the smallest detail so that a world that is like the actual world in every respect except for the placement of a single atom is judged to differ from the actual world in a significant respect. However, it is more reasonable to think, on the contrary, that some respects of similarity and difference are irrelevant

to the interpretation of causally relevant counterfactuals. If this is the case, then worlds that do not differ from the actual world in any significant respect of similarity could count as equally close to the actual world. Moreover, Lewis himself admits that the notion of similarity that operates in his semantics for counterfactuals is a technical notion rather than an intuitive notion of the kind reflected in offhand judgements of similarity. As he notes, the counterfactual 'If Nixon had pressed the button, the world would have been very different' can come out true only when intuitive similarities and differences between worlds are judged to be irrelevant to the interpretation of standard, non-backtracking counterfactuals.

For these reasons, I believe, it is advisable to weaken Lewis's centring principle. So I propose to modify the principle so that the smallest sphere in the system of spheres centred on the actual world may contain other worlds besides the actual world. In the operative technical sense of similarity, other worlds can count as having the same degree of similarity to the actual world as the actual world has to itself. This modification still requires that the actual world be in the smallest sphere, signifying that no world is more similar to the actual world than it is to itself. (Suspending this requirement would amount to suspending the rule of Modus Ponens.) Nonetheless, this simple modification invalidates the problematic inferences discussed above. The inference from '*c* obtains' and '*e* obtains' to 'If *c* were to obtain, *e* would obtain' is invalid: the truth of '*c* obtains' and '*e* obtains' does not guarantee the truth of the counterfactual, since '*c* obtains' may be true and '*e* obtains' false in some non-actual world in the smallest sphere of worlds. Equally, it follows from the fact that the smallest sphere may contain more worlds than the actual world that a counterfactual with a true antecedent does not reduce to the corresponding material conditional. Accordingly, the rule Strengthening the Antecedent is invalid equally for counterfactuals with true antecedents as for those with false antecedents.

With the centring principle weakened in this way, the truth conditions for the first counterfactual (i) in the definition of counterfactual dependence are no longer trivially satisfied when the states of affairs *c* and *e* obtain. The truth of the counterfactual 'If *c* were to obtain, *e* would obtain' requires more than the mere truths '*c* obtains' and '*e* obtains': *e* must obtain in all the worlds closest to the actual world in which *c* obtains. But exactly which are these worlds? Which worlds count as the closest or most similar to the actual world?

Before I try to answer this question, I want to mention some work by James Woodward (2006) that bears on these issues. For the purposes of his argument, Woodward adopts Lewis's definition of counterfactual dependence together with his centring principle. However, he observes that when counterfactuals of type (i) have a property he calls insensitivity, we are much more willing to accept the corresponding causal claim. The notion of insensitivity is one introduced by Lewis for other purposes (1986, Postscript C). But in saying that the counterfactual (i) is insensitive, Lewis and Woodward mean that there is a broad range of background conditions B_i that are not too improbable or far-fetched such that the following counterfactual is true:

(iii) If c were to obtain in circumstances B_i different from the actual circumstances, then e would obtain.

Woodward employs some examples of Lewis's to illustrate his claim that the acceptability of a causal claim goes hand-in-hand with the insensitivity of the corresponding counterfactual of type (i). According to Lewis, shooting your victim at point blank through the heart is an example of insensitive causation since the effect of your action is not sensitive to small variations in background conditions. It is possible to imagine circumstances in which such a shooting would not cause death—for example, the victim is in a state-of-the-art hospital with a surgical team ready to implant an artificial heart—but such circumstances are uncommon and seem far-fetched. As an example of a sensitive counterfactual dependence, Lewis describes the following example. He writes a strong letter of recommendation that causes X to get a job he would otherwise not have got, which in turn causes Y to take another job which he would otherwise not have taken, and so on. The letter thus causes X, Y, and many others, who are similarly displaced from jobs they would otherwise have taken, to meet and marry different people and to have children they would not have had in the absence of the letter. Lewis says that his writing the letter of recommendation caused the death of Y's grandchild, but Woodward remarks that this is a case of sensitive causation because the corresponding counterfactual (i) is sensitive: there are ever so many variations on the actual world such that his writing the letter of recommendation for X did not lead to Y's grandchild being born and so dying.

Woodward is able to explain many interesting facts concerning our willingness to accept causal claims in terms of the relative sensitivity of the corresponding counterfactuals of type (i). He does not build the insensitivity requirement into the truth conditions of these counterfactuals. However, his observations about the importance of the insensitivity of such counterfactuals reflect, I think, the fact that these counterfactuals have non-trivial truth conditions involving many possible worlds in addition to the actual world. He says that causation requires the counterfactual (i) to hold in worlds that do not depart too much from the actual world or do not seem too far-fetched, whereas I say causation requires the counterfactual (i) to be true by virtue of the consequent holding in all of many closest antecedent-worlds. I believe we are talking about the same phenomenon.

But which worlds count as the closest or most similar worlds to the actual world? Which worlds do not depart too much from the actual world or do not seem too far-fetched? Unfortunately, Woodward does not provide a systematic account of how to determine these worlds. Nonetheless, we can draw on another source of insight into these matters: to wit, the psychological literature on counterfactual availability, which investigates the circumstances under which people are inclined to entertain counterfactual suppositions. In drawing on this literature, I shall assume that the degree to which people are ready to entertain counterfactual antecedents tells us something about how close or similar they take the corresponding antecedent-worlds to be to the actual world.

The literature on psychological availability is extensive, but I shall draw mostly on the seminal articles by Kahneman & Tversky (1982) and by Kahneman & Miller (1986). In these articles they proposed that mental simulation played an important role in counterfactual judgements, especially in which an event is judged to be close to having happened. In such cases, they noted, people regularly mentally undo the past by mutating abnormal events to more normal states and seldom, if ever, do they mentally undo the past by changing normal or expected occurrences to abnormal ones. For example, in one study Kahneman & Tversky (1982) gave subjects a story describing a fatal road accident, in which a truck ran a red light and crashed into a passing car, killing its occupant, Mr. Jones. Two versions of the story were constructed: one (labelled route) in which Mr Jones left his home at the regular time but took an unusual route home, and the other (labelled time) in which he took the usual route home but left early to do some chores. In 80% of the responses that mentioned either route or time, subjects indicated that they mentally undid the accident by mutating the exceptional or abnormal event and restoring it back to normality. Thus, in the first version of the story, subjects were inclined to entertain counterfactuals about what would have happened if Mr Jones had taken his usual route home, and in the second version of the story they were more inclined to entertain counterfactuals about what would have happened if Mr. Jones had left work at his usual time. Kahneman & Tversky referred to the norm-restoring mutations as downhill changes and the norm-violating mutations as uphill changes to highlight the mental ease and mental effort with which these types of counterfactual simulations are generated. They wrote: 'The preference for downhill changes is perhaps the major rule that mental simulations obey: it embodies the essential constraints that lend realism to counterfactual fantasies' (1982: 205).

As Kahneman & Miller use the term, a normal event or state of affairs is one that is common, expected, and unsurprising, whereas an abnormal event or state of affairs is one that is exceptional, unexpected, and surprising. It is also important to note that they use the terms 'normal' and 'abnormal' in evaluative as well as non-evaluative senses: a normal event is one that conforms to the norms and an abnormal one is one that violates the norms, where the relevant norms can be evaluative or empirical. So, for example, they describe our practice of mentally undoing reality by mutating abnormal occurrences back to normal ones in terms of our altering non-ideal states of affairs into more ideal ones. They quote a study by Read (1987) in which subjects were taught the rules of a simple, two-person game. Subjects were then shown pictures of the player's hands and were asked to assess what the outcome would have been if players' hands had been different. Subjects chose to modify the outcome by strengthening the losing hand rather than weakening the stronger one. Kahneman & Miller propose the hypothesis: 'When an alternative to an event could be produced either by introducing an improvement in some antecedent or by introducing a deterioration, the former will be more available' (1986: 143).

On the basis of these observations, it would be natural to assume that the spheres of similarity used in evaluating a counterfactual are to be understood purely in terms of

normality or conformity with norms: perhaps the spheres (centred on the actual world) are ordered in terms of what is normal at the actual world, with inner spheres containing worlds with more normal states of affairs and outer spheres containing worlds with less normal states of affairs. (For one elaboration of this conjecture, see Menzies, 2004, 2007.) However, this assumption is not quite correct. For one of the most widely attested phenomena concerning our naïve causal judgements is that we typically see the cause as an abnormal or exceptional state of affairs. For example, Kahneman & Miller write:

> A cause must be an event that could easily have been otherwise. In particular, a causal cannot be a default value among the elements that [the event to be explained] had invoked. The rule that a default value cannot be presented as a cause was noted by Hart and Honoré (1985), who observed that the statement 'It was the presence of oxygen that caused the fire' makes sense only if there is reason to view the presence of oxygen as abnormal. (1986: 149)

And many other philosophers and psychologists have noted the same feature of our causal judgements. As just remarked, Hart & Honoré made a similar observation, but they presented it in an especially insightful way:

> 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 . . . 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. (1985: 29)

If these observations are correct, then the conjecture about ordering spheres of similarity in terms of normality cannot be correct, at least given the working hypothesis of this paper that causal judgements are analytically connected to counterfactuals. For some of these counterfactuals will concern norm-violating causes; and if they are to be true, their closest antecedent-worlds cannot be worlds in which only normal events take place. Some account of the similarity relation is needed that will allow for the realization of antecedents about abnormal states of affairs.

I believe that there is much truth in Hart & Honoré's conjecture in the passage above that the paradigm case of a causal judgement is one in which the cause is a human action that is seen as an intervention or intrusion into the normal course; and that other kinds of causal attributions are to be understood by analogy with this

paradigm case. Transferring this insight to counterfactuals of types (i) and (ii), one can infer that the paradigm case is one in which the antecedent of the type (i) counterfactual is an intentional action that represents an intervention or intrusion into the normal course of events; and in which the antecedent of the type (ii) counterfactual is an omission that allows the normal course of events to proceed in the absence of any intervention. If Hart & Honoré's conjecture is correct, we see other cases by analogy with this case. But how do we do this? I suggest that we do this by generalizing the most important intuitive feature of an intentional action, which is that it represents an independent, exogenous causal influence on the course of events. Various researchers (Pearl, 2000; Spirtes, Glymour, & Scheines, 2000; Woodward, 2003) have introduced the technical notion of an intervention in an attempt to capture and generalize this feature of intentional actions. I propose to use the notion of an intervention, without any precise technical definition in mind, to mean simply a source of causal influence that is exogenous to the causal set-up being considered. It is practically unavoidable to appeal to such a notion in providing the semantics of non-backtracking counterfactuals, as is evidenced by the fact that the standard interpretations of such counterfactuals such as Lewis's (1979 [1986]) and Woodward's (2003) appeal to the concept of a miracle or the technical concept of an intervention. Given this informal understanding of an intervention, I shall also say that a counterfactual is a *deviant counterfactual* if its closest antecedent-worlds are ones in which the antecedent is realized by an exogenous intervention; and say that a counterfactual is a *default counterfactual* if its closest antecedent-worlds are ones in which the antecedent is realized in the normal course of events in the absence of any intervention.

Finally, I am in a position to provide a characterization of the worlds that count as the closest worlds to the actual world for the purposes of evaluating a counterfactual either of type (i) or (ii).

Truth Conditions for Causally Relevant Counterfactuals
Consider a counterfactual 'If P were the case, Q would be the case' where P is about states of affairs at a time t. Consider all those worlds w such that:

(a) P is true at w;
(b) w is exactly like the actual world at all times before a transition period beginning shortly before t;
(c) w conforms to the norms of the actual world at all times after t; and
(d) during the transition period w departs from the norms of the actual world no more than it must to permit P to hold.

The counterfactual is true if and only if Q holds in every such world w.

If the counterfactual is a deviant counterfactual, the worlds w are ones in which P is realized by an exogenous intervention; and if the counterfactual is a default counterfactual, the worlds w are ones in which P is not realized by an exogenous intervention.

Given these truth conditions, I can state my hypothesis concerning the truth conditions for causation in terms of counterfactuals:

Truth Conditions for Causation
A state of affairs *c* causes a wholly distinct state of affairs *e* if and only if (i) if *c* were to obtain, *e* would obtain; and (ii) if *c* were not to obtain, *e* would not obtain, where (i) is a deviant counterfactual and (ii) is a default counterfactual.

These truth conditions are intended to capture the idea that the causation involves two contrasting conditions: the condition represented by the deviant counterfactual (i), according to which *e* obtains in all the worlds in which the course of events proceeds normally after an exogenous intervention brings it about that *c* obtains; and the condition represented by the default counterfactual (ii), according to which *e* does not obtain in all the worlds in which the course of events proceeds normally in the absence of any exogenous intervention. By requiring that a cause satisfy these two counterfactual conditions, I am in effect requiring that a cause be like an intervention that makes a difference to the normal course of events, in line with Hart & Honoré's conjecture.

In motivating the account above, it was natural to focus on the paradigm case in which the abnormal state of affairs considered as a putative cause is a positive intentional action that represents an intervention or intrusion into the normal course of events. However, it is important to keep in mind that the abnormal state of affairs considered as a putative cause may be an omission (or, more generally, an absence). Omissions are typically considered as causes 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 with vaccines; rain by the use of umbrellas. When some harm occurs in violation of 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 the 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. At any rate, the relevant point to keep in mind is that where an omission is being considered as a putative cause, the corresponding type (i) counterfactual, construed as a deviant counterfactual, will describe how an intervention that realizes the omission affects the normal course of events and the type (ii) counterfactual, construed as a default counterfactual, will describe how certain pre-existing practices or procedures would evolve normally in the absence of any such intervention.

There are many controversial features of the proposed truth conditions for counterfactual and causal judgements. But, perhaps, the most controversial is that they allow a role not only for empirical norms, but also evaluative norms, including social, legal, and even moral norms. Many philosophers will regard it as implausible that such norms can influence our counterfactual and causal judgements. Recent empirical work by Chris-

topher Hitchcock & Joshua Knobe (2009), however, demonstrates that such norms do affect our causal judgements. In one experiment they presented subjects with a vignette in which a receptionist in the philosophy department keeps her desk stocked with pens. In contrast to administrative assistants who are permitted to take pens, faculty members are supposed to buy their own. One day, an administrative assistant and Professor Smith walk past the receptionist's desk and both take pens. Later that day, the receptionist needs to take an important message but cannot do so because there are no pens on her desk. When subjects are asked whether the administrative assistant or Professor Smith caused the problem, most subjects say that Professor Smith was the cause. Hitchcock & Knobe conclude that the norm forbidding faculty members from taking the pens led subjects to view Professor Smith's action, but not the administrative assistant's, as the cause of the problem. They constructed several other experiments to rule out the possibility that subjects were conflating questions of causation and blame.

On the account that I wish to propose, our causal judgements about this example are crucially influenced by whether the corresponding counterfactuals of type (i) are true deviant counterfactuals. For the corresponding counterfactuals of type (ii) are clearly both true counterfactuals: it is true that the problem would not have arisen if Professor Smith had not taken a pen and it would not have arisen if the administrative assistant had not taken a pen. The difference in causal status between Professor Smith's action and the administrative assistant's action boils down to the fact that the type (i) counterfactual regarding Professor Smith is a true deviant counterfactual and the one concerning the administrative assistant's action is not. The counterfactual 'If Professor Smith were to take a pen, there would be a problem' is a true deviant counterfactual because in the closest worlds in which Professor Smith takes the pen (in a way that counts as an exogenous intervention), the norms permit the administrative assistant to take a pen and she does so, with the result that a problem arises later when the receptionist searches for a pen. In contrast, the corresponding type (i) counterfactual about the administrative assistant, 'If the administrative assistant were to take the pen, the problem would arise', is not a true deviant counterfactual. In the closest antecedent-worlds in which she takes a pen (in a way that counts as an exogenous intervention), the norms proscribe Professor Smith's taking a pen, with the result that no problem later arises for the receptionist. What differentiates Professor Smith's action from the administrative assistant's is not the fact, taken by itself, that the first is abnormal and the second is normal, but rather that this difference implies that the type (i) counterfactual about Professor Smith is a true deviant counterfactual but the equivalent type (i) counterfactual about the administrative assistant is not.

5 Explanation of the Counterexamples

In this section I shall revisit the counterexamples to Lewis's counterfactual theory of causation, discussed in Section 3. My aim is to apply the account of causation in terms of difference-making counterfactuals to explain our intuitions about these examples.

The first three counterexamples show that Lewis's theory does not distinguish between causes and enabling conditions. My hypothesis about the distinction is that an enabling condition has to satisfy only the counterfactual condition (ii), with the counterfactual typically construed as a deviant counterfactual, whereas a cause has to satisfy both conditions (i) and (ii), with the counterfactual in (i) construed as a deviant counterfactual and the counterfactual in (ii) as a default counterfactual.

Birth and Death

Recall that in this example we judge a car accident to be a cause of a person's death, whereas we do not judge his birth to be so. It is possible to see that the car accident, but not the person's birth, counts as a cause of the person's death, according to the truth conditions above. Since the important difference between these events concerns the status of the type (i) counterfactuals rather than the type (ii) counterfactuals, I shall focus on the former. To see that the type (i) counterfactual, 'If the car accident were to occur, the person would die', is a true deviant counterfactual, perform the following thought experiment: consider the actual world up until a short time before the actual car accident, then imagine that the car accident occurs, not as a result of the actual course of history but as a result of an intervention, and ask whether the person will die in this new course of events as it evolves. Clearly, the answer is that the person will die. On the other hand, a comparable thought experiment shows that the equivalent type (i) counterfactual about the person's birth is a false deviant counterfactual: considering the actual world up until a short time before the person's birth, then imagining that his birth occurs, not as result of the actual course of history but as a result of an intervention, we need not conclude that the person will die. The actual sequence of events from the person's birth to his death involves too many coincidental events and human choices to be replicated in every possible world in which his birth is brought about by an exogenous intervention. Given this failure of the type (i) counterfactual to come out as true, the person's birth does not meet a required condition to be a cause of the person's death. (It is simple to see that the type (ii) counterfactual about the birth is a true deviant counterfactual and so satisfies the requirement to be an enabling condition of the death.) It is tempting to think that the causal difference between the car accident and the person's birth is due to the fact that the former is an abnormal event whereas the latter is part of the normal conditions of the causal set-up. However, the real source of the causal difference lies, I believe, in the more objective fact that the former represents a means by which an agent with suitable powers could bring about a person's death, whereas the latter does not.

The Absence of a Meteor Strike

Recall that in this example we judge the state of affairs of my not being hit by a meteor to be an enabling condition, but not a cause, of my writing this essay. The explanation of our causal judgements about this example follows the pattern of the explanation of the first example. My not being hit by a meteor is not a cause of my writing this paper

because the corresponding type (i) counterfactual, 'If I were not struck by a meteor, I would write this paper', is not a true deviant counterfactual. To see this, consider the actual world up until shortly before I start writing the paper and then imagine that an exogenous intervention ensures that no meteor strikes me during the course of the paper writing. It does not follow that I will write the paper in these imagined circumstances, since my writing of this paper depends on a convergence of many other events—some of which are coincidental and some of which are due to human choices—that are unlikely to be replicated in these imagined circumstances. In contrast, the corresponding type (ii) counterfactual, 'If I were hit by a meteor, I would not write this paper', is a true deviant counterfactual, suggesting that the absence of a meteor strike is an enabling condition of my writing the paper.

Woodward (2006: 25) makes an interesting observation about this example in his discussion of the insensitivity of type (i) counterfactuals. He notes that if I am, in fact, hit by a meteor and die before writing this essay, our causal intuitions change, as we now judge that my being hit by the meteor is indeed a cause of my not writing the essay. But this change in the imagined circumstances of the example involves reversing the counterfactuals of types (i) and (ii): in the new circumstances the type (i) counterfactual is 'If I were hit by a meteor, I would not write this essay', and the type (ii) counterfactual is 'If I were not hit by a meteor, I would write this essay'. The new type (i) counterfactual is a true deviant counterfactual, while the type (ii) counterfactual is not so clearly true as a default counterfactual. As Woodward points out, the truth of the type (i) counterfactual has more weight in our causal judgements than the truth of the type (ii) counterfactual.

The Pot Plant

In this example, we causally discriminate between the gardener's failure to water a pot plant and Queen of England's failure to do so with regards to the death of the pot plant: the gardener's failure counts as a cause but the Queen's failure does not. The explanation of our intuitions about this example is very much like the explanation previously offered of the example involving Professor Smith and the administrative assistant, except that in this case the candidate causes involve omissions rather actions. Once more the difference in causal status between the omissions is due to the difference in the truth value of type (i) counterfactuals, understood as deviant counterfactuals. The type (i) counterfactual 'If the gardener were to fail to water the pot plant, it would die' is a true deviant counterfactual, since in all the worlds in which gardener refrains from performing his watering duties (in a way that counts as an intervention), it is normal for the Queen and other people not to water the plant, with the result that the plant dies in these worlds. In this sense, the gardener's failure is a state of affairs that makes a difference to the normal course of events. In contrast, in all the possible worlds in which the Queen refrains from watering the plant (in a way that qualifies as an intervention) it is normal for the gardener to do so, with the result that the plant survives in these worlds. Here what is crucial for discriminating between the gardener's

failure to water the plant and the Queen's failure is not the fact, taken by itself, that the first is abnormal and the second is not. It is rather that this difference implies that the type (i) counterfactual about the gardener is a true deviant counterfactual while the type (i) counterfactual about the Queen is not.

The next two examples were introduced by Sartorio to illustrate that Lewis's counterfactual theory of causation is afflicted by some Prince of Wales problems to the extent that it generates unwanted positive and negative causes. These examples raise slightly different issues from the examples just considered.

Unwanted Positive Causes

Recall that in this example the Prince of Wales is asked to water a plant but fails to do so, preferring to eat oaten biscuits. However, his priorities are such that if he had not eaten oaten biscuits, he would have watered the plant. Sartorio claims that Lewis's theory implies that both the Prince's failure to water the plant and his eating oaten biscuits count as causes of the plant's death, though intuition dignifies only the first of these as a cause.

The account of causation in terms of difference-making counterfactuals disagrees with Sartorio's causal intuitions in this case. It does agree that the Prince's failure to water the plant is a cause of its death since the type (i) counterfactual 'If the Prince were not to water the plant, it would die' is a true deviant counterfactual and the type (ii) counterfactual 'If the Prince were to water the plant, it would not die' is a true default counterfactual. Surprisingly, however, the account also delivers the result that the Prince's eating oaten biscuits qualifies as an additional cause, since both the corresponding counterfactuals of type (i) and (ii) are true counterfactuals of the right type. The type (i) counterfactual is 'If the Prince were to eat oaten biscuits, the plant would die'. This can be seen to be a true deviant counterfactual, since it follows by a modified Transitivity Rule[4] from the true deviant counterfactual 'If the Prince were to eat oaten biscuits, he would not water the plant' (a truth following from the stipulation that the Prince cannot eat oaten biscuits and water the plant at the same time) and the true deviant counterfactual 'If the Prince were to eat oaten biscuits and not water the plant, it would die'. The type (ii) counterfactual is 'If the Prince were not to eat oaten biscuits, the plant would not die'. This can be seen to be a true default counterfactual since it follows by the modified Transitivity Rule from the true default counterfactual 'If the Prince were not to eat oaten biscuits, he would water the plant' (a stipulated

[4] The rule of inference Transitivity is not valid for counterfactuals: it does not follow from 'If *P* were the case, *Q* would be the case' and 'If *Q* were the case, *R* would be the case' that 'If *P* were the case, *R* would be the case'. However, the following modified rule is valid: from the premises 'If *P* were the case, *Q* would be the case' and 'If *P* and *Q* were the case, *R* would be the case' to the conclusion 'If *P* were the case, *R* would be the case'. (See Lewis, 1973a).

truth about the Prince's preferences) and the true default counterfactual 'If the Prince were not to eat oaten biscuits but water the plant instead, it would not die'.

This result may appear problematic since it contradicts Sartorio's intuition that it is inappropriate to say that the Prince's eating oaten biscuits is a cause of the plant's death. But I suggest that her intuition that the Prince's failure to water the plant differs in causal status from his eating oaten biscuits merely reflects a residual sense that his omission is more abnormal than his eating oaten biscuits. It is easily seen, however, that they cannot differ in causal status since the stipulations of the example ensure that they are counterfactually equivalent. These stipulations ensure that if the Prince were to water the plant, he would not eat oaten biscuits; and that if he were not to eat oaten biscuits, he would water the plant. Given their counterfactual equivalence, it is very plausible to identify them as being one and the same state of affairs.

Unwanted Negative Causes

Recall that in this example the Prince eats so many oaten biscuits that he gets a stomach ache. Sartorio points out that Lewis's theory implies the untoward result that whenever the Prince's eating oaten biscuits precludes a possible action, the omission of that action counts as a cause of his stomach ache. So, for example, the Prince's eating oaten biscuits precludes his watering the plant. But if he had watered the plant, he would not have eaten the oaten biscuits and so would not have got the stomach ache. So Lewis's theory counts his failure to water the plant as a cause of his stomach ache. For similar reasons, a plethora of other omissions such as his failure to talk to the Duke, to walk in his garden, and so on, count as causes of his stomach ache. These negative causes are, in Sartorio's opinion, superfluous and unwanted.

This time the present account delivers verdicts in agreement with Sartorio's. The Prince's eating many oaten biscuits is a cause of his stomach ache since the corresponding types (i) and (ii) counterfactuals are true counterfactuals of the right kind. However, his failure to water the plant does not qualify as a cause since the type (i) counterfactual 'If the Prince were to fail to water the plant, he would get a stomach ache' is false, when interpreted as a deviant counterfactual. If an intervention ensured that the Prince failed to water the plant, it would not follow that he would get a stomach ache, since there are other things the Prince could do instead of watering the plant that are consistent with not getting a stomach ache. In contrast to the previous example, no argument using the modified Transitivity Rule can be deployed to justify this counterfactual. In order to mount such an argument it would have to be argued that the counterfactuals 'If the Prince were to fail to water the plant, he would eat many oaten biscuits' and 'If he were to fail to water the plant and were to eat many oaten biscuits, he would get a stomach ache' are true deviant counterfactuals that follow from the stipulations of the example. While description of the example entails that the second of these is a true deviant counterfactual, it does not entail that the first is. At most the stipulations of the example supply the converse to this first counterfactual: if the Prince were to eat many oaten biscuits, he would fail to water the plant. Consequently, the present account blocks the Prince's failure to water the

plant from counting as a cause of his stomach ache. It is fit and proper that it should do so, since the recipe Sartorio describes would generate an over-abundance of other negative causes for this event, such as his failure to talk to the Duke, his failure to walk in his garden, and so on.

6 Conclusion

Of course, the success of the account of causation in terms of difference-making counterfactuals in dealing with these few examples does not prove that it is correct. For a start, it would need to be shown that the account is able to provide a natural treatment of pre-emption examples that have proved to be such a bane for counterfactual theories. On the other hand, the enrichment of the notion of counterfactual dependence and the definition of causation in terms of this enriched notion, proposed in this paper, are intuitive and psychologically plausible, in my view. I hope that the modest success of this account of causation in explaining the distinction between causes and enabling conditions and the distinction between positive and negative causes might encourage philosophers and psychologists to revisit their views that causal and counterfactuals judgements are not analytically related and that causal and counterfactual reasoning are quite distinct forms of reasoning.[5]

References

Byrne, R. (2005) *The Rational Imagination*. Cambridge, MA: MIT Press.

Collins, J., Hall, N., & Paul, L.(2004) (eds) *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Fine, K. (1975) 'Review of Lewis's *Counterfactuals*', *Mind* 84: 451–8.

Hall, N. (2004) 'Two Concepts of Causation', in Collins, Hall, & Paul (2004), pp. 225–76.

Hart, H. & Honoré, A. (1985) *Causation in the Law* (2nd edition). Oxford: Clarendon Press.

Hitchcock, C. & Knobe, J. (2009) 'Cause and Norms', *Journal of Philosophy* 106: 587–612.

Kahneman, D. & Miller, D. (1986) 'Norm Theory: Comparing Reality to its Alternatives', *Psychological Review* 80: 136–53.

—— Slovic, P., & Tversky, A. (1982) (eds) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

—— & Tversky, A. (1982) 'The Simulation Heuristic', in Kahneman, Slovic, & Tversky (1982), pp. 201–10.

Knowles, D. (1990) (ed.) *Explanation and its Limits*. Cambridge: Cambridge University Press.

Lewis, D. (1973a) *Counterfactuals*. Oxford: Basil Blackwell.

—— (1973b) 'Causation', *Journal of Philosophy* 70: 556–67. Reprinted with Postscripts in Lewis (1986).

[5] I would like to thank the participants at the 2007 Workshop on Causal and Counterfactual Understanding at the University of Warwick for comments that have decisively influenced the contents of this paper.

—— (1979) 'Counterfactual Dependence and Time's Arrow', *Nous* 13: 455–76. Reprinted in Lewis (1986). Page references to reprinted version.

—— (1986) *Philosophical Papers, Volume II*. Oxford: Oxford University Press.

—— (2000) 'Causation as Influence', *Journal of Philosophy* 97: 182–97. Reprinted in Collins, Hall, & Paul (2004). Page references to the reprinted version.

Mackie, J. (1974) *The Cement of the Universe*. Oxford: Clarendon Press.

Mandel, D. (2003) 'Judgment Dissociation Theory: An Analysis of Differences in Causal, Counterfactual, and Covariational Reasoning', *Journal of Experimental Psychology: General* 132: 419–34.

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

—— (2004) 'Difference-making in Context', in Collins, Hall, & Paul (2004), pp. 139 –80.

—— (2007) 'Causation in Context', in Price & Corry (2007), pp. 191–223.

Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

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

Read, S. (1987) 'Constructing Causal Scenarios: A Knowledge Structure Approach to Causal Reasoning', *Journal of Personality and Social Psychology* 52: 288–302.

Sartorio, C. (2009) 'The Prince of Wales Problem for Counterfactual Theories of Causation', in A. Hazlett (ed.) *New Waves in Metaphysics*, pp. 259–76. Basingstoke: Palgrave Macmillan.

Schaffer, J. (2000) 'Trumping Preemption', *Journal of Philosophy* 97: 165–81.

Spirtes, P., Glymour, C., & Scheines, R. (2000) *Causation, Prediction, and Search* (2nd edition). Cambridge, MA: MIT Press.

Woodward, J. (1990) 'Supervenience and Singular Causal Statements', in Knowles (1990), pp. 215–26.

—— (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

—— (2006) 'Sensitive and Insensitive Causation', *Philosophical Review* 115: 1–50.

10

Counterfactual and Causal Thoughts about Exceptional Events

Ruth M.J. Byrne[1]

1 Counterfactuals, Causes, and Exceptions

Suppose you are in a car accident on the way home from work. It is likely that you will spend some time afterwards thinking about the accident. When something unexpected or exceptional happens, it is natural to try to understand it by thinking about what led to it, and about how it could have turned out differently. You may try to work out the causes of the accident, for example, 'the collision was caused by the other car coming around the corner too far over on my side of the road'. You might also think about how the situation could have turned out differently, for example 'the collision could have been avoided if only I had swerved out of the way when the other car came around the corner'. Philosophers have suggested that the concept of causation relies on the contrast between reality and a counterfactual alternative (e.g. Hume, 1739/1978; Lewis, 1973; Mackie, 1974; Mill, 1843/1967). Psychologists have identified close links between counterfactual thoughts and causal attributions (e.g. Hilton, 1990; McGill & Klein, 1993; Roese & Olson, 1995). However, psychologists have also discovered that causal and counterfactual thoughts diverge in important ways (e.g. Mandel & Lehman, 1996; N'gbala & Branscombe, 1995). In this chapter I wish to consider several strands of evidence that shed light on the complex relationship between people's counterfactual thoughts and their causal thoughts.

In this chapter I focus on exceptional or unexpected antecedents and outcomes (Byrne, 2005). People are prompted to construct causal explanations and to create counterfactual alternatives very often after an unexpected outcome, such as a car accident (e.g. Klauer & Migulla, 1995; Roese & Olson, 1995; Sanna & Turley,

[1] *Address for correspondence:* School of Psychology and Institute of Neuroscience, Lloyd Building, Trinity College Dublin, University of Dublin, Dublin 2, Ireland. Email: rmbyrne@tcd.ie.

1996). They may do so because an unexpected outcome indicates a violation of their understanding of the usual causal relationships. Likewise, they tend to focus on exceptional antecedents rather than normal ones when they think about how things could have turned out differently 'if only' (e.g. Gavanski & Wells, 1989; Kahneman & Tversky, 1982a). Consider the following scenario (based on Kahneman & Tversky, 1982a: 204):

> Paul was 47-years-old, the father of three and a successful banker. His wife had been ill at home for several months. On the day of the accident, Paul left his office at the regular time. He sometimes left early to take care of household chores at his wife's request, but this was not necessary on that day. Paul did not drive home by his regular route. The day was exceptionally clear and Paul told his friends at the office that he would drive along the shore to enjoy the view. The accident occurred at a major intersection. The light turned amber as Paul approached. Witnesses noted that he braked hard to stop at the crossing, although he could easily have gone through. His family recognised this as a common occurrence in Paul's driving. As he began to cross after the light changed, a truck charged through the intersection at top speed, and rammed Paul's car from the left. Paul was killed instantly. It was later ascertained that the truck was driven by a teenage boy, who was under the influence of drugs. As commonly happens in such situations, Paul's family and their friends often thought and often said 'if only . . . ' during the days that followed the accident. How did they continue this thought?

When people create a counterfactual alternative to reality, they tend not to change the most improbable event—two cars being in exactly the same place at exactly the same time. Most people tend to imagine the same alternative to reality. They tend to say, 'If only he had gone home by his usual route' (Kahneman & Tversky, 1982a). People mentally 'undo' the exceptional event to make it normal, and they do so regardless of which event in the scenario is indicated to be exceptional. For example, some participants in the experiment received a version in which they were told instead (Kahneman & Tversky, 1982a: 204):

> On the day of the accident, Paul left the office earlier than usual, to attend to some household chores at his wife's request. He drove home along his regular route. Paul occasionally chose to drive along the shore, to enjoy the view on exceptionally clear days, but that day was just average.

People mentally change the exceptional event, regardless of whether it is the route or the time. The participants who received the information that Paul had taken an unusual route home tended to say, 'if only he had gone home by his usual route', and participants who received the information that Paul had gone home at an unusual time said, 'if only he had gone home at his usual time' (Kahneman & Tversky, 1982a).

In this chapter I focus on the counterfactual thoughts and causal explanations that people create for exceptional and unexpected antecedents and outcomes (Byrne, 2005). I consider four questions that have puzzled psychologists and philosophers interested in causation and counterfactuality: (1) How do counterfactual 'if only' thoughts affect causal thoughts? (2) How do causal thoughts affect counterfactual thoughts? (3) How do causal and counterfactual thoughts differ? And (4) how do

semi-factual 'even if' thoughts affect causal thoughts? I would like to suggest that the answers to these four questions follow from two central claims. First, counterfactual thoughts require people to construct a mental representation of two possibilities, the counterfactual conjecture and the presupposed factual reality. As a result, counterfactual thoughts make explicit some knowledge, including causal knowledge, that is otherwise implicit. Second, there are different sorts of causes, that is, there are causes with different logical meanings which have different psychological representations. 'Strong' causes are necessary and sufficient for their outcome, and they can be mentally represented economically by thinking initially about just one of their possibilities; 'enabling' causes are necessary but not sufficient for their outcome, and they must be mentally represented initially by thinking about at least two of their possibilities. As a result, counterfactual thoughts often change enabling causes.

2 Question 1: How Do Counterfactual Thoughts Affect Causal Thoughts?

One view of how people identify a cause is that they understand that 'A caused B' by thinking about 'If A hadn't happened then B wouldn't have happened'. Consider the following conjecture: 'What if, in the summer of 1941, Hitler had chosen to make his major attack not into Soviet Russia but across the Eastern Mediterranean, into Syria and the Lebanon? Would he have avoided the defeat he suffered outside Moscow that winter?' (Keegan, 1999). Historical analyses can benefit from rigorous counterfactual construction, because such counterfactual thoughts can affect the understanding of the causal relationships between events (e.g. Tetlock & Lebow, 2001).

Children's counterfactual thoughts can also influence their causal thoughts (e.g. Harris, 2000; Robinson & Beck, 1999). For example, consider a scenario in which a doll named Carol walks across a clean white floor with dirty shoes and makes the floor dirty with her footprints. Children as young as 3-years-old can answer counterfactual questions such as, 'What if Carol had taken her shoes off, would the floor be dirty now?' 3- and 4-year-olds seem to be able to engage in counterfactual thinking to reach a causal conclusion (Harris, 2000; Harris, German, & Mills, 2000).

Counterfactual thoughts also increase adults' judgements of causality. In one experiment, participants considered a scenario in which a taxi driver refused to give a couple a lift and they had to take their own car instead. On the way, a bridge collapsed as they crossed over it and they were killed (Gavanski & Wells, 1989). Some participants were told that the taxi driver had earlier crossed the bridge safely, and others were told that the taxi driver was also killed crossing the bridge. Participants in the first condition, who could presumably create the counterfactual alternative, 'if only the taxi driver had given the couple a lift, they would still be alive', tended to judge the taxi driver's role as being more causal in the couple's death, compared to participants in the second condition, who could presumably create the semi-factual alternative, 'even if the taxi

driver had given the couple a lift, they would still have died' (Gavanski & Wells, 1989). How do counterfactual thoughts help people work out the causes of events? The account I wish to advance here is that counterfactual thoughts make explicit some information that would otherwise remain implicit in an individual's understanding of the causal relationship between events (Byrne, 2005). The next section provides a brief sketch of this view and some evidence that corroborates it.

Counterfactuals and Possibilities

A counterfactual such as 'if the taxi driver had given the couple a lift, they would have made it across the bridge' requires individuals to think about alternative possibilities (Byrne, 2005). They must envisage the counterfactual conjecture, 'the taxi driver gave the couple a lift and they made it across the bridge' and they must also think about the known or presupposed facts, 'the taxi driver did not give the couple a lift and they did not make it across the bridge'. They also need to keep track of the epistemic status of the two possibilities (Johnson-Laird & Byrne, 1991).

Counterfactual conditionals are special (e.g. Lewis, 1973; Stalnaker, 1968). Compare the counterfactual conditional in the subjunctive mood, 'if the taxi driver *had given* the couple a lift, they *would have* made it across the bridge', to a factual conditional in the indicative mood, 'if the taxi driver gave the couple a lift, they made it across the bridge'. The two conditionals seem to mean something very different, and their difference is captured in the different possibilities that people envisage when they understand them.

We suggest that people understand conditionals by imagining possibilities (Johnson-Laird & Byrne, 2002). On this view they construct 'mental models', that is, mental representations that correspond to the way the world would be if the assertion were true (Johnson-Laird, 1983). The possibilities that people keep in mind are guided by a principle of truth: they generally do not tend to think about false possibilities (Johnson-Laird & Byrne, 2002). For example, the indicative conditional corresponds to several true possibilities including: 'the taxi driver gave the couple a lift and they made it across the bridge', and 'the taxi driver did not give the couple a lift and they did not make it across the bridge'. But it rules out as false the possibility 'the taxi driver gave the couple a lift and they did not make it across the bridge'. We propose that people do not tend to think about this false possibility when they understand the conditional.

On this view, the possibilities that people envisage are also guided by a principle of parsimony. People generally do not tend to think about all of the true possibilities, perhaps because of the limitations of working memory (Johnson-Laird, Byrne, & Schaeken, 1992). Instead, they tend to think initially about a single, true possibility when they understand the indicative conditional, e.g. 'the taxi driver gave the couple a lift and they made it across the bridge'. They can 'flesh out' their models to include the other true alternative possibilities if need be. There is considerable evidence that people find it easier to think about a single possibility than about multiple possibilities (Johnson-Laird, 2006). For example, when they make inferences that require them

to envisage multiple models, they make more errors and take longer than when the inference requires them to envisage a single model (e.g. Johnson-Laird, Byrne & Schaeken, 1992).

In contrast, for the counterfactual conditional, 'if the taxi driver *had given* the couple a lift, they *would have* made it across the bridge', people think of two possibilities from the outset, the conjecture, and the presupposed facts. Counterfactuals are psychologically intriguing: people are able to think readily about what is false, temporarily supposed to be true. They must keep track of the epistemic status of these alternative possibilities; one corresponds to a counterfactual conjecture, and the other corresponds to the presupposed or known facts. Their initial mental representation is richer for the counterfactual conditional than the indicative one (see Byrne & Johnson-Laird, 2009, for a review).

We have found considerable evidence to support the view that people think of two possibilities when they understand a counterfactual conditional and only a single possibility when they understand a factual conditional. For example, we discovered that people read a conjunction 'there were no roses and there were no lilies' faster when they have first read the counterfactual 'if there had been roses then there would have been lilies' rather than its indicative counterpart. In one experiment, Carlos Santamaria, Orlando Espino, and I gave 32 participants from the University of La Laguna a set of simple stories to read (Santamaria, Espino, & Byrne, 2005). The stories were about various different topics; for example, one story concerned the flowers that were available in a florist's shop that a boy and his sister visited. Each story contained either a counterfactual conditional, such as 'if there had been roses then there would have been lilies', or a factual conditional, such as 'if there were roses then there were lilies'. Later in the story the participants read one of various different conjunctions about the events, e.g. 'there were roses and there were lilies' or 'there were no roses and there were no lilies'. Each participant read 48 simple stories in a within-participant design, consisting of 6 instances of each different type of story (containing a factual or counterfactual conditional, and one of four sorts of conjunctions). Participants read each sentence of the story on a computer screen and pressed a key to see the next sentence. We recorded how much time they spent reading each sentence.

The results showed that participants read the conjunction 'not A and not B' 114 ms faster when it was primed by a counterfactual compared with a factual conditional and the difference was statistically significant. They took the same length of time to read the 'A and B' conjunction whether it was primed by a counterfactual or a factual (only a 56 ms difference). There were no differences in the times people took to read 'not A and B' or 'A and not B' (see Santamaria, Espino, & Byrne, 2005, Experiment 1).

We have also discovered that people make different inferences from counterfactual conditionals compared to indicative ones. People are willing to endorse different inferences from a counterfactual conditional compared to a factual conditional because of the differences in their mental representations. Alessandra Tasso and I found that people make the inferences that correspond to the 'not-A and not-B' possibility more

readily from the counterfactual than from the factual conditional. We gave participants inferences of the following sort (based on Byrne & Tasso, 1999: 730):

Imagine you are given information about the location of different people in different places. You know that: *If Linda had been in Dublin then Cathy would have been in Galway*. Then you are told, *Cathy was not in Galway*. What conclusion, if anything, follows from these premises?

In one experiment, we compared such inferences to similar ones based on a factual conditional, e.g. 'If Linda was in Dublin then Cathy was in Galway'. The factual conditional is represented in an initial set of models that contains a single true possibility: 'Linda is in Dublin and Cathy is in Galway' and so the inference is a difficult one to make. But the counterfactual conditional is represented in a set of models that captures the two possibilities, the counterfactual conjecture: 'Linda is in Dublin and Cathy is in Galway', and the presupposed facts: 'Linda is not in Dublin and Cathy is not in Galway', and so we predicted that the inference should be made more readily from the counterfactual conditional.

We tested 130 participants from the University of Padua. The experiment contained four different sorts of conditionals as a between-participants variable (present and past indicative conditionals and present and past subjunctive conditionals), and it contained as a within-participants variable four different inferences, and participants carried out one instance of each of them. The results show that when participants were given the information that Cathy was not in Galway, which matches information they have represented explicitly in their model of the presupposed facts for a counterfactual conditional, they concluded that Linda was not in Dublin. They did so more often from the counterfactual than from the factual conditional, 66% versus 42%, and the difference was significant (Byrne & Tasso, 1999). The evidence that people consider two possibilities when they understand a counterfactual is difficult for theories of conditional reasoning based on the computation of probabilities to explain; it rules out the idea that mental representations follow a 'singularity principle' by which individuals mentally represent only a single possibility (e.g. Evans & Over, 2004; see also Oaksford & Chater, 2007).

The evidence sheds light on a possible answer to the question of how counterfactual 'if only' thoughts affect causal thoughts. The answer requires a consideration of the mental representations of counterfactuals. On our account, counterfactual thoughts require participants to construct a mental representation of two possibilities, the counterfactual conjecture and the presupposed factual reality. Counterfactual thoughts may thus make explicit some knowledge that is otherwise implicit (Byrne, 2005). When reasoners understand a counterfactual conditional such as 'if she hadn't studied then she wouldn't have passed the examination' they think about the conjecture, and they mentally represent the components mentioned explicitly in the conditional, 'she didn't study and she didn't pass the examination'. But they also think about the presupposed facts, and they mentally represent information that is implicit in the conditional, 'she studied and she passed the examination'. When this implicit information is made explicit in their mental representation of the two possibilities, it helps to

make clearer the causal relationship between studying and passing the examination. As a result people may learn more from creating counterfactual alternatives than from generating causal explanations (see also Kahneman, 1995; Roese, 1994).

3 Question 2: How Do Causal Thoughts Affect Counterfactual Thoughts?

Causal thoughts can influence counterfactual thoughts. For example, experts in military and political history who are committed to the causal principle that nuclear deterrence works tend to be more dismissive of counterfactual conjectures such as 'if Kennedy had listened to his hawkish advisers, the Cuban missile crisis would have become nuclear' compared to experts who are not committed to the causal principle (Tetlock & Lebow, 2001).

How do causal thoughts affect counterfactual thoughts? The account I wish to consider here is based on the idea that different causal relations have different logical meanings, which require people to keep in mind different sorts of possibilities (Johnson-Laird & Byrne, 1991). We propose that strong causes require people to envisage just a single possibility for them. As a result they place fewer demands on cognitive capacity than counterfactuals; people do not need to inhibit their view of what actually happened to envisage an alternative conjecture (Beck, Riggs, & Gorniak, 2009; Beck, Robinson, Carroll, & Apperly, 2006). The next section provides a brief sketch of this view and some evidence that corroborates it.

Causal Relations and Possibilities

The meaning of a strong causal relation, for example, 'Heating water to 100 degrees causes it to boil', identifies the cause as necessary and sufficient for the outcome. It is logically consistent with the possibility 'The water was heated to 100 degrees and it boiled' and a second possibility 'The water was not heated to 100 degrees and it did not boil'. Psychologically, people tend to mentally represent this sort of equivalence or 'biconditional' causal relation by thinking initially about just the first possibility (e.g. Goldvarg & Johnson-Laird, 2001). Importantly, on this account there are different sorts of causes—strong causes, weak causes, enabling causes—and we will return to consider some of them in the next section.

Because people tend to envisage a single possibility when they think about strong causes—whereas they think about two possibilities for counterfactual alternatives—it is often easier to generate causal thoughts than counterfactual thoughts. Thinking about alternative possibilities is difficult even for adults: people take longer and they make more errors when they must do so (e.g. Johnson-Laird & Byrne, 1991).

We discovered some evidence to support the view that causal explanations are easier to create than counterfactual alternatives when we observed individuals' spontaneous thoughts. Consider the following scenario (based on McEleney & Byrne, 2006: 241):

You're moving house to start a new job in a different city. The night before you leave, you write down your thoughts about the move in your diary: ... I've got mixed feelings about moving to a place where I know hardly anyone—my friends and social life are so important to me. But I'm sure it will be easy to settle in to the new town—I've never had any trouble making new friends ...

A lot happens in your first two weeks in the new town. During your first week at work, a staff dinner is held. You have to go because your boss has asked all the staff to be there. You enjoy the evening and meet a lot of people. That weekend, your next-door neighbours invite you to a party. Most of the people who live on your road will be there. However, that evening you're extremely ill with the flu, so you can't go. The next week, you happen to bump into an old friend who lives in the town and he insists on showing you around. He takes you out the following evening and he introduces you to a lot of his friends. A few days later, a colleague tells you there's a membership vacancy at her sports club. You think joining would be a good way to meet people, but there's no way you can afford the membership fee.

Six weeks after the move, things have turned out nothing like you had expected. You haven't made any real friends in the new town and you feel very lonely and isolated. You are very upset, and very surprised.

In one experiment Alice McEleney and I gave 248 participants from Trinity College Dublin different versions of this scenario which either contained a good outcome or a bad one (e.g. you made no real friends), and which was either expected or unexpected (e.g. at the outset you are sure it will be easy to settle in to the new town—you've never had any trouble making new friends, but in the end things have turned out nothing like you expected). The four antecedents that led to the outcome were either the result of a controllable decision (e.g. you decide to go to the cinema instead of the party) or were the result of an uncontrollable event (e.g. you are ill and so you cannot go to the party).

Most importantly, participants were not directed to produce causal or counterfactual thoughts. They were simply asked to write a diary entry (McEleney & Byrne, 2006: 241):

That night, while thinking over the events of the past six weeks, you realize that you haven't written in your diary since the night before you moved. In the space below, jot down what you might write in your diary about your current situation and the events leading up to it.

We considered a counterfactual thought to be any change to an event that would change the outcome, e.g. 'If only I had gone to that party, I would have made friends'. A causal explanation was any statement that attempted to explain why the outcome occurred, e.g., 'I haven't made friends because I didn't go to that party'.

The results showed that the participants spontaneously produced more causal thoughts than counterfactual thoughts. The 248 participants produced 197 counterfactual thoughts overall—an average of 0.79 per participant—and 402 causal explanations overall—an average of 1.62 per participant, and the difference was statistically significant (see McEleney & Byrne, 2006, Experiment 1). They spontaneously created more counterfactual thoughts when they thought about a bad outcome rather than a good one (M = 1.22 vs M = 0.38), and likewise, they produced more causal explanations when they thought about a bad outcome rather than a good one

($M = 1.91$ vs $M = 1.34$, and both differences were statistically significant). However, the experiment also showed important differences between causal and counterfactual thoughts and we will return to those in the next section.

Causal thoughts can affect counterfactual thoughts. Part of the answer to the question about how they do so comes from a consideration of their mental representation. Strong causal relations concern causes that are necessary and sufficient for their outcome, and so people can mentally represent them economically by envisaging just one of their possibilities. As a result, causal thoughts about strong causal relations are easier to mentally represent and maintain than counterfactual thoughts, which require people to envisage two possibilities. Consistent with this proposal, the evidence from people's spontaneous counterfactual and causal thoughts shows that they produce more causal explanations than counterfactual thoughts (McEleney & Byrne, 2006). As it is often easier to generate causal thoughts—at least about strong causal relations—than counterfactual thoughts, causal thoughts can often precede counterfactual thoughts and so have the opportunity to direct subsequent counterfactual thoughts.

4 Question 3: How Do Causal and Counterfactual Thoughts Differ?

Causal and counterfactual thoughts can diverge quite dramatically. For example, in one experiment participants were given a scenario in which an accident occurs when the actor was driving home on a route that was unusual for him; a drunk driver swerved across the road and crashed into him (Mandel & Lehman, 1996). When participants were asked the cause of the accident, they said it was the drunk driver; but when they were asked how the accident could have turned out differently, they said 'if only the actor had driven home by a different route' (Mandel & Lehman, 1996). The distinction is also observed in the thoughts of individuals who have been involved in accidents, who continue to mull over how they could have prevented the accident, even though they are aware that it was not their fault (Davis et al. 1995).

One view is that counterfactual thoughts focus on controllable antecedents, whereas causal thoughts focus on antecedents that covary with an outcome (Mandel & Lehman, 1996). Another view is that counterfactual thoughts focus on necessary causes, whereas causal thoughts focus on sufficient causes (N'gbala & Branscombe, 1995). Of central importance to our views about the differences between counterfactual and causal thoughts is the identification of different sorts of causal relations (Johnson-Laird & Byrne, 1991; Goldvarg & Johnson-Laird, 2001). Causes can differ both in their logical meaning and in their psychological representation. Consider an *enabling* cause, for example, 'turning the ignition key causes the car to start'. The cause is necessary but not sufficient for the outcome. Its meaning is logically consistent not only with the same two possibilities as a strong cause, 'the ignition key was turned and the car started' and 'the ignition key was not turned and the car did not start', but also with a third

possibility, 'the ignition key was turned but the car did not start'—because the car starting depends on other things, for example it must have petrol. As a result, its psychological representation is also different from a strong cause: people must think from the outset about more than a single possibility in order to capture the meaning of the enabling causal relation (Byrne, 2007).

When people identify a cause they must distinguish between causes, enablers, and irrelevant factors, e.g. the colour of the pot the water is heated in (e.g. Cheng & Novick, 1991). A missing enabler can prevent the outcome, e.g. 'there was no petrol in the tank, which prevented the car from starting' (Byrne, 1989, 2005), as can a preventing factor or disabler, e.g. 'diesel in the tank prevented the car from starting' (Cummins, 1995; see also Adler & Rips, 2008; Sloman, 2005).

We propose that counterfactual thoughts focus on enabling causes (Byrne, 2005), perhaps because people envisage the same sorts of possibilities for counterfactuals and enablers. The next sections consider three strands of evidence which shed light on these differences between causal and counterfactual thoughts, especially for exceptional events. The evidence suggests that causal explanations tend to focus on strong causal relations and counterfactual thoughts focus on enabling causes. The first strand of evidence in the next section comes from the experiment described earlier about spontaneous counterfactual and causal thoughts.

Spontaneous Causal and Counterfactual Thoughts about Exceptional Outcomes and Antecedents

In the experiment described earlier with Alice McEleney in which we examined people's spontaneous causal and counterfactual thoughts as they wrote a diary entry about their move to a new town and making new friends there, we found intriguing differences in the spontaneous causal and counterfactual thoughts that participants created. For example, they created more counterfactual thoughts when they thought about controllable events rather than uncontrollable ones ($M = 1.1$ vs $M = 0.5$, and the difference was statistically significant); in contrast there was no effect of controllability on the frequency of the causal explanations that they provided (see McEleney & Byrne, 2006). Perhaps more strikingly, the participants produced more causal explanations when they thought about an unexpected outcome compared to an expected one ($M = 1.79$ vs $M = 1.47$, and the difference was statistically significant); there was no effect on the frequency of their counterfactual thoughts for expected and unexpected outcomes. The result shows that people produce as many *spontaneous* counterfactual thoughts about normal outcomes as about exceptional ones, whereas they produce more spontaneous causal explanations about exceptional outcomes than normal ones.

In a subsequent experiment we showed that the finding holds whether it is the outcome that is unexpected, as the previous experiment showed, or the antecedents that are exceptional. We gave 84 participants from Trinity College Dublin scenarios of the same sort as the one above about moving to a new town, but this time they all had

uncontrollable, unexpected, bad outcomes. Half of the participants read scenarios in which each of the four antecedent events was described as exceptional (it rarely happens) and the other half read scenarios in which each antecedent was described as normal (it usually happens) (see McEleney & Byrne, 2006, Experiment 2).

The results showed once again that participants produced more causal explanations than counterfactual thoughts. Participants produced more causal thoughts than counterfactual thoughts when the antecedents were exceptional ($M = 1.62$ vs $M = 0.36$, and the difference was statistically significant) and when they were normal ($M = 1.00$ vs $M = 0.55$, and the difference was statistically significant). Most strikingly, they produced as many *spontaneous* counterfactual thoughts for exceptional antecedents as they did for normal antecedents ($M = 0.55$ vs $M = 0.36$). The result runs counter to the observation that people tend to change exceptional events in their counterfactual thoughts. As we saw in the example at the outset of the chapter, when participants are prompted by a scenario, for example, about an individual who was in a car accident when they took an unusual route home, they tend to focus on the exceptional event, e.g. 'If only he had gone home by his usual route' (Kahneman & Tversky, 1982a). But we have discovered that when they create spontaneous counterfactual thoughts about an unexpected bad outcome, they focus equally on exceptional and normal antecedents (McEleney & Byrne, 2006).

In contrast, people produced more causal explanations for exceptional antecedents than for normal ones ($M = 1.62$ vs $M = 1.00$, and the difference was statistically significant). The evidence demonstrates an important difference between spontaneous causal and counterfactual thoughts: people produce more spontaneous causal explanations for unexpected exceptional events whereas they produce as many spontaneous counterfactual thoughts about unexpected, exceptional events as about expected, normal ones.

The Goal-driven Nature of Counterfactual Thoughts about Exceptional Events

Even in scenario-based counterfactual thoughts about exceptional antecedents, such as the situation in which an individual was in a car accident when they took an unusual route home, the tendency for people to change exceptional events to be normal is diminished when they know of a good reason for the exceptional action (Zhang, Bonnefon, & Deng, 2004). For example, when Paul is in a car accident and he took an exceptional route home because he had to collect medicine from the chemist for his sick wife, people do not wish 'if only he had taken his usual route'. Likewise, when people think about how an exceptional outcome might have turned out differently, they change exceptional actions that led to it; but when they think about how a normal outcome might have turned out differently, they change both normal and exceptional antecedents (Gavanski & Wells, 1989).

Counterfactual thoughts appear to be guided by the goal of changing the outcome (Byrne, 2002; Roese, Sanna, & Galinsky, 2005). James Dixon and I have discovered that people change exceptional actions to be normal only when they believe that the

normal action would have led to a better outcome, e.g. 'if Paul had driven home by his regular route, he would not have been in a car accident'. We have found that they change exceptional actions to be *even more* exceptional when they believe the more exceptional action would have led to a better outcome and the normal one would not.

Consider a card game in which an actor must choose to place a large bet, a medium bet, or a small bet (Dixon & Byrne, 2011). The actor plays against three other players: the small-bet player only matches small bets, the medium-bet player matches small or medium bets, and the large-bet player matches small, medium, or large bets. The actor usually chooses to place a small bet. However, in this exceptional instance, the actor chooses to place the medium bet. As a result, the medium-bet player and the large-bet player both match his bet. It turns out that the actor had better cards than the large-bet player, but the medium-bet player had the best cards and so the medium-bet player won. Imagine the actor thinks 'things would have been different if . . . '. How do you think he would complete this thought?

The counterfactual alternative created by changing the actor's exceptional action (placing the medium bet) to be like his usual action (placing the small bet) would not change the outcome to be better (if he had placed his usual small bet he still would have lost, because the medium bet player would have matched his bet and won). But the counterfactual alternative created by changing the actor's exceptional action to be like the even more exceptional action (placing the large bet) *would* change the outcome to be better (if he had placed a large bet he would have won, because the medium-bet player would not have played).

In one experiment James Dixon and I gave 60 participants from Trinity College Dublin the card game scenario. Their 'if only' thoughts changed the exceptional bet to be even more exceptional rather than to be normal (67% vs 10%, and the difference was statistically significant). They did so even when they were first given a question designed to encourage them to think strategically—'explain what you think is the best decision for the actor' (57% vs 17%, and the difference was statistically significant).

Of course, we also showed that when the usual action would lead to a better outcome, participants undid the exceptional event to be normal in this sort of scenario. We gave a different group of 60 participants from Trinity College Dublin a version of the card scenario in which the large-bet player had better cards than both the actor and the medium-bet player, and so the actor lost. Counterfactual thoughts that focus on the other exceptional bet cannot result in a win outcome in this scenario—if the actor had placed a large bet he still would have lost—whereas those that focus on the usual action at least result in a smaller loss—if he had placed his usual small bet he would have lost less money. In this situation their 'if only' thoughts undid the exceptional bet to be normal rather than exceptional (43% vs 10%, $p < .012$), even when they first thought strategically (67% vs 17%, and the differences were statistically significant).

The discovery highlights an important aspect of counterfactual thinking about exceptional actions: people change exceptional actions to be normal only when they are optimal, that is, they believe that the normal action would have led to a better

outcome; they change exceptional actions to be *even more* exceptional when they believe the more exceptional action would have led to a better outcome and the normal one would not.

Reasons and Actions Versus Causes and Effects

When people generate counterfactual thoughts about a causal sequence, they tend to imagine alternatives to causes more than to their effects (Wells, Taylor, & Turtle, 1987). When participants were given a scenario in which a man arrived too late to buy a stereo in a sale because he was delayed by four causally linked events, they tended to think 'if only' about the first event more than the other three (Wells et al. 1987; Segura, Fernandez-Berrocal, & Byrne, 2002). Just as events tend to have causes, so people tend to act for a reason (Zeelenberg, van den Bos, van Dijk, & Pieters, 2002; see also Byrne & McEleney, 2000). But there is philosophical debate about whether reasons are like causes (e.g., Donnellan, 1967; Smith & Jones, 1986). The relationship between some reasons and actions may not be as straightforward as the relationship between a strong cause and its effect, for example, the one-to-one mapping of a strong cause and its effect contrasts with the many-to-one mapping for some sorts of reasons and actions or the one-to-many mapping for other sorts of reasons and actions.

In one experiment Clare Walsh and I examined whether or not people think 'if only' more often about a reason that led to an action, just as they tend to think 'if only' more often about a cause that led to an effect. Consider the following scenario, from Walsh & Byrne (2007: 481):

Sam is an avid tennis player. Recently, he took part in an important tournament in France. On Sunday he was due to take part in the semi-finals. Normally, Sam does some training the day before an important match. *However, on Saturday he decided to rest instead. He wanted to be fresh for the following day.* As a result, he didn't do any training that day. The day of the match was very hot. Although Sam had often played in hot conditions, he really didn't like it. That morning a draw was held to see which of the other three semi-finalists Sam would play. After the draw, Sam made a choice of two rackets, one made of synthetic material and one made of metal. He finally decided to use the metal racket because it was most suited to his opponent's manner of play. That afternoon, Sam played the match and lost. He was very disappointed.

We gave 36 participants from Trinity College Dublin six different scenarios of this sort in a within-participants design and we gave them the following task (from Walsh & Byrne, 2007: 467):

As commonly happens in such situations, Sam often thought, 'if only ...' during the days and weeks that followed. Please list in the order of importance, four ways in which he may have completed this thought so that the outcome would be different.

We compared scenarios that contained an action and its reason, as in the scenario above, 'However on Saturday he decided to rest instead (action). He wanted to be fresh for the following day (reason)', to scenarios that contained an event and its cause,

'However, on Saturday the courts were closed (event). Heavy rain the night before meant that the courts were waterlogged (cause)'. Our primary interest was in the comparison between these two sorts of sequence, reason-action and cause-effect, but we also examined cause-action sequences, 'However, on Saturday he decided to rest instead (action). After the long journey, he was unable to stay awake (cause)', and reason-effect sequences, 'However, on Saturday the courts were closed (event). This was to save the lawn for the upcoming matches (reason)'. We compared them to control scenarios, 'However, on Saturday he decided to rest instead. He went to his room', and 'However, on Saturday the courts were closed. He went to his room' (see Walsh & Byrne, 2007, Experiment 1).

The results showed that for reason-action sequences, participants imagined an alternative to the action more than the reason (83% vs 14%, and the difference was statistically significant). For cause-effect sequences, they imagined an alternative to the effect or to the cause equally (56% and 33%). Overall, people thought 'if only' about an action more than an effect (78% vs 59%, and the difference was statistically significant), when it followed a reason (83% vs 56%) or a cause (64% vs 56%). Complementarily, people thought 'if only' about a cause more than a reason (38% vs 11%), whether it led to an action (42% vs 14%) or to an effect (33% vs 8%, and both differences were statistically significant).

The evidence demonstrates an important difference in counterfactual thoughts about causal relations and intentional, or reason-based, relations: in a cause-effect sequence, people think 'if only' more often about the cause than about its effect, but in a reason-action sequence, they think 'if only' more often about the action than its reason.

How do causal and counterfactual thoughts differ? The three strands of evidence considered in this section shed further light on the differences between them for exceptional events: people produce more spontaneous causal explanations for exceptional events than normal ones whereas they produce as many spontaneous counterfactual thoughts about exceptional events as about normal ones; people change exceptional actions to be normal only when they believe that the normal action would have led to a better outcome; they change exceptional actions to be *even more* exceptional when they believe the more exceptional action would have led to a better outcome and the normal one would not; and people think 'if only' more often about a cause than about its effect, but they think 'if only' more often about an action than its reason.

The results are consistent with the idea that counterfactual and causal thoughts diverge in part because causal thoughts tend to focus on strong causes, whereas counterfactual thoughts tend to focus on enabling causes. In our view, there are different sorts of causes and they are mentally represented differently. Strong causes are necessary and sufficient for their outcomes and so they can be represented psychologically by thinking initially about just a single possibility. Enabling causes are necessary but not sufficient for their outcomes and so they must be mentally represented initially by thinking about at least two of their possibilities. The representation

of counterfactual thoughts, which requires the consideration of two alternative possibilities, and the mental representation of enabling causes is thus similar (Byrne, 2005).

5 Question 4: How Do Semi-Factual 'Even If' Thoughts Affect Causal Thoughts?

Philosophers distinguish between counterfactual 'if only' assertions and semi-factual 'even if' assertions (e.g., Barker, 1991; Goodman, 1973). Counterfactual 'if only' thinking has been hypothesized to play a key role in establishing causation but very little is known about the effects of semi-factual 'even if' thinking on causal judgements (e.g. Boninger Gleicher, & Strathman, 1994; Sherman & McConnell, 1996; Branscombe et al. 1996).

Rachel McCloy and I carried out an experiment to examine the effects of semi-factual thoughts on causal judgements (McCloy & Byrne, 2002). We adapted a scenario from Boninger et al. (1994) and gave it to 367 participants from Trinity College Dublin (from McCloy & Byrne, 2002: 49–50):

You are a runner and since the age of eight you have competed in the sprint races in local track and field events. Up through school you had won every race in which you had competed. It was at the age of 13 that you began to dream about the Olympics. At the age of 18, before starting college, you decide to give the Olympics one all out shot. You make the Irish Olympic team for the 400 metre race. On the day before the 400 metre race, in a freak accident during training, you sprain your left ankle. Although there is no break or fracture, when you try to run, the pain is excruciating.

Your trainer tells you about many advances in pain killing medications and assures you that you will still be able to participate. He recommends that you choose between two drugs, both legal according to Olympic guidelines. One is a well-known pain killer that has been proved effective but also has some serious side effects including temporary nausea and drowsiness. The other pain killer is a newer and less well-known drug. Although the research suggests that the newer drug might be a more effective pain killer, its side effects are not yet known because it has not been widely used. After considerable thought, you elect to go with the more well-known drug.

On the day of the race, although there is no pain in your ankle, you already begin to feel the nausea and find yourself fighting off fatigue. You finish in fourth place, only 1 tenth of a second from a bronze medal, 4 tenths from a silver, and 5 tenths from a gold medal.

After the event, you learn that some athletes in other events who were suffering from similar injuries used the other, newer drug. They felt *no* pain but experienced *the same* side effects.

Imagine that in the days and weeks following the race you thought 'even if . . . '. How would you complete this thought?

Some participants read a final paragraph describing a same-outcome alternative as outlined above, and a second group of participants read an alternative-outcome version:

After the event, you learn that some athletes in other events who were suffering from similar injuries used the other, newer drug. They felt *no* pain and experienced *no* side effects.

For a third group of participants this paragraph was omitted, and they received no information about other athletes' experiences with the newer drug.

In each of these three groups one-third of the participants were asked to imagine that in the days and weeks following the race they thought 'even if . . . ' and their task was to complete this thought; another third of the participants were asked to imagine that they thought 'if only . . . ' and to complete this thought; and the remainder were given no sentence completion task. All the participants were asked to put themselves in the place of the runner and to make the causal judgement: 'To what extent do you think your decision to take the well-known drug led to your failure to obtain an Olympic medal in the 400 metre race?'. They gave their answer on a 9-point scale where 1 indicated that they believed their decision was not at all causal, and 9 indicated that they believed it was the most important cause of their failure.

The results showed that 'even if ' thoughts *decreased* participants' ratings of the causal role of the decision (to take the more well-known drug) compared to no thoughts ($M = 4.82$ vs 5.67, and the difference was statistically significant). 'If only' thoughts did not increase or decrease ratings of the causal role of the decision compared to no thoughts ($M = 5.23$ vs 5.67). Likewise, the results showed that participants' ratings of causality *decreased* when the same outcome occurred for the other drug compared to when they were given no information about the outcome from the other drug ($M = 4.26$ vs 5.25); and their ratings of causality *increased* when a different outcome occurred for the other drug compared to when they were given no information about the outcome from the other drug ($M = 6.23$ vs 5.25, and both differences were statistically significant).

How do semi-factual thoughts affect people's view of the causes of events? The account I wish to advance here is that semi-factual thoughts, like counterfactual thoughts, make explicit some information that would otherwise remain implicit in an individual's understanding of the causal relationship between events (Byrne, 2005). The next section provides a brief sketch of this view and some evidence that corroborates it.

Semi-factual Conditionals and Possibilities

People understand a semi-factual assertion, such as 'even if there had been roses there would have been lilies' by thinking about the conjecture 'there were roses and there were lilies', and also by thinking about some presupposed facts, in this case 'there were no roses and there were lilies' (Byrne, 2005). They keep track of the epistemic status of the two possibilities as imagined or real.

We have discovered considerable evidence to support the view that people think of these two possibilities when they understand a semi-factual conditional. In an experiment similar to the one outlined earlier, Carlos Santamaria, Orlando Espino, and I gave a different set of 32 participants a set of simple stories to read (Santamaria, Espino, & Byrne, 2005). The stories contained either a semi-factual conditional, 'even if there had been roses there would have been lilies' or a factual conditional, 'if there were roses there were lilies'. Later in the story the participants read a conjunction of the events, e.g., 'there were roses and there were lilies' or 'there were no roses and there were lilies'. We again recorded how much time they spent reading each sentence.

The results showed that participants read the conjunction 'not A and B' 104 ms faster when it was primed by a semi-factual compared with a factual conditional, and the difference was statistically significant. They took the same amount of time to read the 'A and B' conjunction whether it was primed by a semi-factual or a factual conditional (only a 32-ms difference). There were no differences in the times people took to read 'not A and not B' or 'A and not B' (Santamaria, Espino, & Byrne, 2005).

We have also discovered that people are willing to endorse different inferences from a semi-factual conditional compared to a factual conditional because of these differences in their mental representations. Sergio Moreno-Rios, Juan Garcia-Madruga, and I have found that people make the inferences that correspond to the 'not-A and B' possibility more readily from the semi-factual than from the factual conditional. In one experiment, we gave 32 participants from the University of Granada inferences of the following sort (based on Moreno-Rios, Garcia-Madruga, & Byrne, 2008: 201):

Even if Luis had pressed the button, the machine would have started. Luis did not press the button. Therefore . . . (1) The machine started. (2) The machine did not start. (3) There is no valid conclusion.

We compared them to similar inferences based on the factual conditional 'If Luis pressed the button, the machine started'. The factual conditional is represented in an initial set of models that correspond to a single true possibility: 'Luis pressed the button and the machine started'. The information in the second premise 'Luis did not press the button' does not match with any of the information in the single possibility that reasoners have thought about from the outset and so participants often say that nothing follows. If reasoners think about a second alternative, 'Luis did not press the button and the machine did not start', they endorse the conclusion 'the machine did not start'. But if they think also about a third possibility, 'Luis did not press the button and the machine started anyway', they will resist the inference, which is a fallacy on this conditional interpretation.

The semi-factual conditional is represented differently from the factual conditional, in a set of models that captures two possibilities, the conjecture: 'Luis pressed the button and the machine started' and the presupposed facts: 'Luis did not press the button and the machine started'. The information in the second premise in the problem above, 'Luis did not press the button' matches the information in this second possibility. As a result we predicted that the inference 'the machine did not start' would not be made from the semi-factual conditional. In fact, we predicted that participants would instead make an asymmetrical inference: given 'Luis did not press the button', they would tend to infer 'the machine started'.

The experiment had a within-participants design. Each participant received two conditionals (past indicative 'if' and past subjunctive 'even if'), and four different sorts of inferences. They carried out four instances of each inference and so there were 32 trials, each with a different content. The results showed that when participants were

given the information that Luis did not press the button, which matches information they have represented explicitly in their model of the presupposed facts for the semi-factual conditional, they made fewer inferences to the conclusion that the machine did not start from the semi-factual than from the factual conditional (41% vs 88%) and instead they made the asymmetrical inference to 'the machine started' (44% vs 7%, and both differences were statistically significant) (see Moreno-Rios, Garica-Madruga, & Byrne, 2008, Experiment 1).

The answer to the question of how semi-factual 'even if' thoughts affect causal thoughts requires a consideration of the mental representation of semi-factual thoughts. When reasoners understand a semi-factual conditional such as 'even if he had been rich she wouldn't have married him' they think about the conjecture, and they mentally represent the components mentioned explicitly in the conditional, 'he was rich and she didn't marry him'. But they also think about the presupposed facts, and they mentally represent information that is implicit in the conditional, 'he wasn't rich and she didn't marry him'. When this implicit information is made explicit in their mental representation of the two possibilities, it helps to make clearer the denial of a causal relationship between his being rich and her marrying him. Semi-factual thoughts require participants to construct a mental representation of two possibilities, the semi-factual conjecture and the presupposed factual reality. 'Even if' thoughts thus make explicit some knowledge about the absence of a causal relation that is otherwise implicit (Byrne, 2005).

6 Conclusions

In this chapter I have considered four questions that have been of interest to psychologists and philosophers interested in the relation between counterfactual and causal thoughts, with special reference to exceptional events. (1) The answer to the first question—how do counterfactual 'if only' thoughts affect causal thoughts?—depends on the observation that counterfactual thoughts require participants to construct a mental representation of two possibilities, the counterfactual conjecture and the presupposed factual reality (Byrne, 2005). Evidence from priming studies and from inference studies corroborates this proposal (Santamaria et al. 2005; Byrne & Tasso, 1999). Counterfactual thoughts may thus make explicit some knowledge that is otherwise implicit (Byrne, 2005). (2) The answer to the second question—how do causal thoughts affect counterfactual thoughts?—is that people mentally represent strong causes by envisaging a single possibility, whereas they represent counterfactual alternatives by envisaging two possibilities. Causal thoughts are thus easier to mentally represent than counterfactual thoughts. The evidence from individuals' spontaneous counterfactual and causal thoughts shows that people produce more causal explanations than counterfactual thoughts, about unexpected outcomes (McEleney & Byrne, 2006). (3) The answer to the third question—how do causal and counterfactual thoughts differ?—rests on the proposal that there are different sorts of causes and

they are mentally represented differently (e.g. Johnson-Laird, 2006). Strong causes are represented by thinking initially about a single possibility, whereas enabling causes are represented initially by thinking about two possibilities. The representation of counterfactual thoughts and enabling causes is similar and hence counterfactual thoughts focus on changes to enabling causes (Byrne, 2005). Three strands of evidence illustrate further some of the divergences of causal and counterfactual thoughts: people generate more spontaneous causal explanations for exceptional antecedents than for normal ones but as many spontaneous counterfactual thoughts for both of them (McEleney & Byrne, 2006). Exceptional actions are undone to be even more exceptional rather than to be normal to bring about a better outcome (Dixon & Byrne, 2011). And in cause-effect sequences, causes are changed more than effects, but in reason-action sequences, actions are changed more than reasons (Walsh & Byrne, 2007). (4) The answer to the fourth question—how do semi-factual 'even if' thoughts affect causal thoughts?—depends on the observation that semi-factual thoughts require participants to construct a mental representation of two possibilities, the semi-factual conjecture and the presupposed factual reality (Byrne, 2005). Once again, evidence from priming studies and from inference studies corroborates this proposal (Santamaria et al. 2005; Moreno-Rios, Garcia-Madruga, & Byrne, 2008). Semi-factual thoughts, like counterfactual thoughts, may make explicit causal knowledge that is otherwise implicit. The causal explanations that people construct and the counterfactual alternatives that they create for exceptional and unexpected antecedents and outcomes sheds some light on the complex relationship between causal and counterfactual thoughts.

References

Adler, J.E. & Rips, L.J. (2008) (eds) *Reasoning: Studies of Human Inference and its Foundations*. Cambridge: Cambridge University Press.

Barker, S. (1991) 'Even, still and counterfactuals', *Linguistics and Philosophy* 14: 1–38.

Beck, S.R., Riggs, K.J., & Gorniak, S.L. (2009) 'Relating developments in children's counterfactual thinking and executive functions', *Thinking and Reasoning* 15: 337–54.

—— Robinson, E.J., Carroll, D.J., & Apperly, I.A. (2006) 'Children's thinking about counterfactuals and future hypotheticals as possibilities', *Child Development* 77(2): 413–26.

Boninger, D.S., Gleicher, F., & Strathman, A. (1994) 'Counterfactual thinking: From what might have been to what may be', *Journal of Personality and Social Psychology* 67: 297–307.

Branscombe, N.R., Owen, S., Garstka, T.A., & Coleman, J. (1996) 'Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome?', *Journal of Applied Social Psychology* 26: 1042–67.

Byrne, R.M.J. (1989) 'Suppressing valid inferences with conditionals', *Cognition* 31: 61–83.

—— (1997) 'Cognitive processes in counterfactual thinking about what might have been', *The Psychology of Learning and Motivation* 37: 105–54.

—— (2002) 'Mental models and counterfactual thinking', *Trends in Cognitive Sciences* 6: 405–45.

—— (2005) *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA.: MIT Press.

—— (2007) 'Precis of *The Rational Imagination: How People Create Alternatives to Reality*', *Behavioral and Brain Sciences* 30: 439–80.

—— & Johnson-Laird, P.N. (2009) '"If" and the problems of conditional reasoning', *Trends in Cognitive Sciences* 13: 282–7.

—— & McEleney, A. (2000) 'Counterfactual thinking about actions and failures to act', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26: 1318–31.

—— Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000) 'The temporality effect in counterfactual thinking about what might have been', *Memory & Cognition* 28: 264–81.

—— & Tasso, A. (1999) 'Deductive reasoning with factual, possible, and counterfactual conditionals', *Memory & Cognition* 27: 726–40.

Cheng, P.W. & Novick, L.R. (1991) 'Causes versus enabling conditions',*Cognition* 40: 83–120.

Cummins, D.D. (1995) 'Naïve theories and causal deduction', *Memory and Cognition* 23: 646–58.

Davis, C.G., Lehman, D.R., Wortman, C.B., Silver, R.C., & Thompson, S.C. (1995) 'The undoing of traumatic life events', *Personality and Social Psychology Bulletin* 21: 109–24.

Dixon, J. & Byrne, R.M.J. (2011) 'Counterfactual thinking about exceptional actions'. *Memory & Cognition*, in press.

Donnellan, K.S. (1967) 'Reasons and causes'. In: *The Encyclopedia of Philosophy*, P. Edwards (ed.), vol 17, pp. 85–8. New York: The Macmillan Company.

Espino, O., Santamaria, C., & Byrne, R.M.J. (2009) 'People think about what is true for conditionals, not what is false: Only true possibilities prime the comprehension of "if"', *Quarterly Journal of Experimental Psychology* 62: 1072–8.

Evans, J.S.B.T. & Over, D.E. (2004). *If*. Oxford: Oxford University Press.

Gavanski, I. & Wells, G.L. (1989) 'Counterfactual processing of normal and exceptional events', *Journal of Experimental Social Psychology* 25: 314–25.

Goldvarg, E. & Johnson-Laird, P.N. (2001) 'Naïve causality: a mental model theory of causal meaning and reasoning', *Cognitive Science* 25(4): 565–610.

Goodman, N. (1973). *Fact, fiction and forecast (3rd edition)*. Cambridge, MA: Harvard University Press.

Harris, P.L. (2000) *The Work of the Imagination*. Oxford: Blackwell.

—— German, T., & Mills, P. (1996) 'Children's use of counterfactual thinking in causal reasoning', *Cognition* 61: 233–59.

Hilton, D.J. (1990) 'Conversational processes and causal explanations', *Psychological Bulletin* 107: 65–81.

Hume, D. (1739/1978) *A Treatise of Human Nature*. Oxford: Oxford University Press.

Johnson-Laird, P.N. (1983) *Mental Models*. Cambridge: Cambridge University Press.

—— (2006) *How We Reason*. Oxford: Oxford University Press.

—— & Byrne, R.M.J. (1991) *Deduction*. Hillsdale: Erlbaum Associates.

Johnson-Laird P.N. & Byrne R.M.J. (2002) 'Conditionals: a theory of meaning, pragmatics, and inference', *Psychological Review* 109: 646–78.

—— —— & Schaeken, W. (1992) 'Propositional reasoning by model', *Psychological Review* 99: 418–39.

Kahneman, D. (1995) 'Varieties of counterfactual thinking'. In: *What Might Have Been: The Social Psychology of Counterfactual Thinking*, N.J. Roese & J.M. Olson (eds), pp. 375–96. Mahwah, New Jersey: Erlbaum.

—— & Miller, D.T. (1986) 'Norm theory: Comparing reality to its alternatives', *Psychological Review* 93: 136–53.

—— & Tversky, A. (1982a) 'The simulation heuristic'. In: *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, & A. Tversky (eds.), pp. 201–11. Cambridge: Cambridge University Press.

—— —— (1982b) 'The psychology of preferences', *Scientific American* 246: 160–73.

Klauer, K.C. & Migulla, G. (1995) 'Spontaneous counterfactual processing', *Zeitschrift für Sozialpsychologie* 26: 34–42.

Keegan, J. (1999) 'How Hitler could have won the war'. In: *What If? Military Historians Imagine What Might Have Been*, R. Cowley (ed.). New York: Penguin Putnam.

Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.

McCloy, R. & Byrne, R.M.J. (2000) 'Counterfactual thinking about controllable actions', *Memory & Cognition* 28: 1071–78.

—— —— (2002) 'Semifactual "Even If" Thinking', *Thinking & Reasoning* 8: 41–67.

McEleney, A. & Byrne, R.M.J. (2006) 'Spontaneous causal and counterfactual thoughts', *Thinking and Reasoning* 12: 235–55.

McGill, A.L. & Klein, J.G. (1993) 'Contrastive and counterfactual thinking in causal judgment', *Journal of Personality and Social Psychology* 64: 897–905.

Mackie, J. (1974) *The Cement of the Universe*. Oxford: Clarendon Press.

Mandel, D.R., Hilton, D.J., & Catellani, P. (2005) *The Psychology of Counterfactual Thinking*. London: Routledge.

—— & Lehman, D.R. (1996) 'Counterfactual thinking and ascriptions of cause and preventability', *Journal of Personality and Social Psychology* 71(3): 450–63.

Mill, J.S. (1843/1967) *A System of Logic: Raciocentive and Inductive*. London: Longmans.

Moreno-Rios, S., Garcia-Madruga, J., & Byrne, R.M.J. (2008) 'Semifactual "even if" Reasoning', *Acta Psychologica* 128: 197–209.

N'gbala, A. & Branscombe, N.R. (1995) 'Mental simulation and causal attribution: When simulating an event does not affect fault assignment', *Journal of Experimental Social Psychology* 31: 139–62.

Oaksford, M. & Chater, N. (2007) *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Rips, L.J. (2009) 'Two causal theories of counterfactual conditionals', *Cognitive Science* 34: 175–221.

Robinson, E.J. & Beck, S. (2000) 'What is difficult about counterfactual reasoning?'. In: *Children's Reasoning and the Mind*, P. Mitchell & K.J. Riggs (eds), pp. 101–19. Hove, UK: Psychology Press.

Roese, N.J. (1994) 'The functional basis of counterfactual thinking', *Journal of Personality and Social Psychology* 66(5): 805–18.

—— & Olson, J.M. (1995) 'Counterfactual thinking: A critical overview'. In: *What Might Have Been: The Social Psychology of Counterfactual Thinking*, N.J. Roese & J.M. Olson (eds). Mahwah, NJ: Erlbaum.

—— Sanna, L.J., & Galinsky, A.D. (2005) 'The Mechanics of Imagination: Automaticity and Control in Counterfactual Thinking'. In: *The New Unconscious*, R.R. Hassin, J.S.Uleman, & J.A. Bargh (eds), pp. 138–70. New York: Oxford University Press.

Sanna, L.J. & Turley, K.J. (1996) 'Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence', *Personality and Social Psychology Bulletin* 22(9): 909–19.

Santamaria, C., Espino, O., & Byrne, R.M.J. (2005) 'Counterfactual and semifactual conditionals prime alternative possibilities', *Journal of Experimental Psychology: Learning, Memory and Cognition* 31: 1149–54.

Segura, S., Fernandez-Berrocal, P., & Byrne, R.M.J. (2002) 'Temporal and causal order effects in counterfactual thinking', *Quarterly Journal of Experimental Psychology* 55: 1295–305.

Sherman, S.J., & McConnell, A.R. (1996) 'The role of counterfactual thinking in reasoning', *Applied Cognitive Psychology* 10: 113–24.

Sloman, S. (2005) *Causal Models: How People Think About the World and Its Alternatives*. Oxford: Oxford University Press.

Smith, P. & Jones, O.R. (1986) *The Philosophy of Mind: An Introduction*. Cambridge: Cambridge University Press.

Stalnaker, R.C. (1968) 'A theory of conditionals'. In: *Studies in Logical Theory*, N. Rescher (ed.), pp. 98–112. Oxford: Basil Blackwell.

Tetlock, P.E. & Lebow, R.N. (2001) 'Poking counterfactual holes in covering laws: cognitive styles and historical reasoning', *American Political Science Review* 95: 829–43.

Thompson, V. & Byrne, R.M.J. (2002) 'Reasoning about things that didn't happen', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28: 1154–70.

Walsh, C.R. & Byrne, R.M.J. (2007) 'The effects of reasons for acting on counterfactual thinking', *Thinking and Reasoning* 13: 461–83.

Wells, G.L., Taylor, B.R., & Turtle, J.W. (1987) 'The undoing of scenarios', *Journal of Personality and Social Psychology* 53: 421–30.

Zeelenberg, M., van den Bos, K., van Dijk, E., & Pieters, R. (2002) 'The inaction effect in the psychology of regret', *Journal of Personality and Social Psychology* 82: 314–27.

Zhang, J., Bonnefon, J.-F., & Deng, C. (2004) 'Regret and reasons to act or not to act'. Unpublished manuscript. Sociology Institute, Shanghai Academy of Social Science.

11

Causation First: Why Causation is Prior to Counterfactuals

Dorothy Edgington

In the first part of this essay I give reasons for thinking that causation is a very basic concept, so basic that the hope is dim of giving an informative account of the reductive sort, '*c* causes *e* iff . . . '. The second part has some criticisms of the attempt to give such an account in terms of counterfactuals. The third part argues that, conversely, to the extent that there is one standard, default way of understanding counterfactuals, we need to appeal to causal notions in saying what it is. In the final part I argue that even if there is one standard, default way of assessing counterfactuals, perfectly proper uses of the counterfactual conditional are not always of that kind, and, in context, a much wider class of counterfactual judgements are permissible. So counterfactuals are, as it were, too promiscuous a kind of judgement to yield an account of causation.

1

In the early part of the twentieth century causation was conspicuously absent from the philosophers' tool kit. Implicitly or explicitly, it was considered to be too problematic a concept for use in the elucidation of other phenomena. For instance, it did not occur to Nelson Goodman (1946) that an appeal to causal notions might help with his attempt to decide which facts are co-tenable with a counterfactual supposition. And C.G. Hempel was dismissive of the idea that causation itself (when not reduced to a statement of law) had any useful role in a theory of explanation: an attempt to explain by citing a cause would be 'like a note saying there is hidden treasure somewhere' (1965: 349), but not saying where. This distrust of the notion of causation derives from the work of David Hume (1739: Book I, Part 3; 1748: sections 4–7) and I will discuss this shortly.

In the 1960s, things changed. There was H.P. Grice's causal theory of perception (1961): what it is to see a vase requires there to be appropriate causal relations between the perceiver and the vase. Donald Davidson argued that reasons for action were causes of the action (1963). Alvin Goldman developed a causal theory of (empirical) knowledge

(1967). In Saul Kripke's theory of reference (1980), using a name correctly requires causal chains connecting the use to the bearer of the name. And David Lewis gave a causal theory of explanation (1986b): to explain an event is to give information about its causal history. I have mentioned some of the main innovators, but the change in perspective was widespread. By and large, the change was not a consequence of any great advance in the theory of causation. Rather, causation came to be seen as a basic concept in terms of which other phenomena could be elucidated.

The exception to this generalization is Lewis, who did give an analysis of causation in terms of counterfactual conditionals, and had a theory of the truth conditions of counterfactual conditionals to boot, so he could, with good analytic conscience, use causation to explain other things.

I don't think Lewis's approach succeeds, as I shall argue below. And I believe the correct perspective takes causation as one of our basic concepts, certainly more basic than counterfactual conditionals. First I shall make some remarks about Hume.

Hume made causality appear to be a very problematic concept because, he argued, it was epistemologically highly problematic: not discoverable by reason; and not 'present to the senses'. Hume's first negative claim is no doubt correct—we do not discover what causes what just by mathematical or logical reasoning, we do so by experience. Nevertheless, there are significant difficulties with the details of his claim. Causation is a relation between things in the world, as Hume seems to allow. It is propositions, or relations between them, that are discovered by reason. The epistemic status of a proposition asserting a causal relation between entities depends on how the entities are described. As Davidson (1967) pointed out, if we refer to an entity, *c*, as the cause of *e*, 'the cause of *e* caused *e*' is (pretty much) knowable a priori.[1]

At first sight, this looks like a trick. We need to distinguish intrinsic descriptions of events as they are in themselves, from extrinsic, relational descriptions, and restrict Hume's thesis to the former. But once one notices that almost all our descriptive vocabulary has causal implications, it is hard to see how to do this, and Davidson's point is a relatively trivial manifestation of a widespread, deep phenomenon.

The things which cause and are caused are, typically, changes in the properties of an object—an object's turning blue, or becoming harder, or increasing in weight. To be blue is to be such as to *cause* (under certain conditions) certain visual sensations; a thing is hard to the extent that it cannot easily be *made* to change its shape, soft to the extent that its parts are easily displaced; to increase in weight is to be more resistant to *efforts to make it move*; and so on. Secondary qualities like *blue* are dispositions to *affect* us in certain ways. Primary qualities, like hardness and weight, are such that to understand them one must (in the words of Gareth Evans) 'master a set of interconnected

[1] 'Pretty much' because it presupposes that some unique thing caused *e*. But this could be fixed by a preceding conditional clause, 'If any one thing caused *e*, then the cause of *e* caused *e*'.

principles which make up an elementary theory—of primitive mechanics—into which the properties fit, and which alone give them their sense . . . One must learn how bodies compete for occupancy of positions of space, of the resistance one body may afford to another, and so on' (Evans, 1985: 269).

A.J Ayer, in *Probability and Evidence*, valiantly trying to defend Hume, is acutely aware of the difficulty of conceiving and describing a 'bedrock of fact' purged of all causal implications. Our ordinary means of describing the world, he allows, is 'over-charged with reference to casual properties' (1972: 115; see also 6–10). To get to the 'purely factual level' we need to subtract from our description of things any logical implications about their powers.

But it is not at all clear that this can be done. A property may just be its potential for contributing to the causal powers of the things that have it; so that if we abstract away from the casual consequences of a thing, there is nothing left. The levels at which we describe the world, if this is the case, are causal all the way down. And if this is right, there is no lower level of fact to which causality can be reduced.[2]

Even if we do not go this extreme, the fact remains that our ordinary ways of describing the world, the kinds of things it contains and the properties they have, is shot through with causal implications. We do learn about causality by experience, but virtually all we learn about the world involves causality.

Hume raised two epistemological difficulties about causation, the first, there being no logical relations between cause and effect, meaning you cannot deduce causes from effects. The above considerations are relevant to this. Logical relations do not relate objects, but descriptions of objects, and virtually all our descriptions of objects have causal implications. His second difficulty was that one doesn't perceive causality. Again, if the above points are correct—if the way we describe the world is causal through and through—a strong case can be made for the perception of a good deal of causality too. As Elizabeth Anscombe says, 'As surely as we learned to call people by name and to report from seeing it that the cat is on the table, we also learned to report from seeing it that someone drunk up the milk, or that the dog made a funny noise, or that things were cut or broken by the things that cut and broke them ... [Hume] confidently challenged us to "produce an instance wherein the efficacy is plainly discoverable by the mind, and its operations obvious to our consciousness or sensation". Nothing easier: is cutting, is drinking, is purring not efficacy?' (Anscombe, 1971: 69).[3] We could add, do we not feel the force of the wind or the impact of a collision? Are we not aware of bringing about changes as we rearrange the furniture or cover a piece of paper with the marks of a pen?

[2] See Stalnaker, 1984: 157–60; Strawson, 1985: 115–36; Shoemaker, 1980: 114.

[3] Anscombe also remarks that Hume is 'helped, in making his argument that we don't perceive "efficacy", by his curious belief that "efficacy" means much the same thing as "necessary connection" (1969: 69). Her conception of causation, bringing things about, has no such connotation.

Of course we can be mistaken, or deluded, in these observations. But this is true of all perception. Indeed, focusing on mistakes strengthens the case that causality is integral to the character of perceptual experience. Consider this distinction: a visual illusion often persists even when we know it is illusory; the Müller-Lyer lines still look as if they are of different length even when we have measured them and discovered them to be equal. On the other hand, if I make an inference from what I observe, and then discover I was wrong, I simply discard the belief. I came to the conclusion that you were away, having tried to phone you many times, but then discover that your phone was out of order, though it rings normally. The belief that you were away is discarded; no illusion persists.

Now, taking an example from Christopher Peacocke (1985: 156), suppose I'm watching a forklift truck raising a metal box—as it seems. I am then told that the box is actually being pulled up by a magnet, the truck is exerting no force (it is merely a back-up device). The illusion still persists: it look just as if the truck is lifting the box. Another example: a game on a computer screen, on which 'little men' move about and fire shots which, if they hit, eliminate other little men. Actually, we know, what we are watching is the common effect of a lot of electronic circuitry: no causality runs from one 'man' to the other. But the illusion persists: it looks just as if the causality does so run.

Maybe a world which is not causally structured is, in some thin sense, possible, but I do not think there could be knowledge or experience of a world which is not causally structured; for knowledge of the world requires that it impinges on our senses: it leaves its imprint on us. (Even a philosopher such as Leibniz, who denies this, and has in its place 'pre-established harmony', does not deny that there is causality involved, for he has it that God so arranges—so causes—our experience faithfully to reflect the world. It is also significant that even our famous sceptical hypotheses, involving evil demons or brains in vats, still involve causation of our sensations, in non-standard ways. So it seems harder to think away causation from the world than almost anything else.) To the extent that knowledge is a practical ability that manifests itself in action, that is the other side of the causal coin—ourselves impinging on the world. So, it seems to me, causality is a fundamental feature of a knowable world.

2

The counterfactual theory of causation was propounded by Lewis (1973/1986a): where *c* and *e* are two distinct events, *e* depends causally on *c* iff, if *c* had not occurred, *e* would not have occurred. *C* causes *e* iff there is a chain of causally dependent events from *c* to *e*.

At the time of writing Lewis had not given us a theory of events. He filled that lacuna when his paper 'Causation' was reprinted in his *Philosophical Papers*, Volume II (1986a), with an essay, 'Events' (1986c), which made clear that, at least for the purpose of studying causation, 'events' were to be interpreted widely, to include persisting states. Also, it should be noted that the relation he tries to define could be called the 'is a factor in the causal ancestry of' relation. He leaves to pragmatics the question of when it

is appropriate to select one from a set of causally relevant conditions as 'the' cause. Thus, his target is what James Woodward (this volume) calls the 'broad' notion of causation. 'It would be strange, though not I think false, to say in any ordinary context that the availability of petrol was a cause of the crash', Lewis says (1986b: 216).

Lewis begins by discussing some well-known difficulties for regularity theories of causation. The first difficulty for a regularity theory is distinguishing law-like from accidental generalizations. Lewis has his own account of this distinction, so let us grant it for the sake of argument. The version of a regularity theory he discusses is this: *c* causes *e* iff *c*'s occurrence is sufficient, given the laws and other circumstances, for *e*'s occurrence, i.e. there are laws and other truths about the situation from which, together with the proposition that *c* occurs, we can deduce that *e* occurs. The remaining difficulties he sees are:

(a) The problem of effects, that is, of distinguishing causes from effects. If flipping the switch is sufficient, given the laws and circumstances, for the light's going on, so the light's going on may be sufficient, given the laws and the circumstances, for the flipping of the switch. The theory wrongly pronounces that *e* caused *c*.
(b) Epiphenomena: the spots may be sufficient, given the laws and the circumstances, for the subsequent fever; the theory wrongly pronounces the spots the cause of the fever (whereas they are both effects of a common cause).
(c) Pre-emption: the movement of the truck is sufficient, given the laws and the circumstances, for the movement of the box; but, contrary to the theory, it is the magnet which is causing the box to move.

The first problem, of distinguishing causes from effects, could be solved by adding to the analysis that causes are earlier than their effects. But Lewis rightly rejects such a solution: it rules out, a priori, by fiat, serious hypotheses about simultaneous or backward causation and possible causal loops. If such things are impossible, they are not *trivially* impossible. It would preclude a non-trivial explanation of the direction of time as the predominant direction of causation. And it leaves untouched the problem of epiphenomena: the spots are earlier than the fever, and may be sufficient, given the laws and circumstances, for the fever, but they do not cause the fever, they are effects of a common cause. Another way of putting the objection to simply defining causation as unidirectional is this: the interesting question then becomes why there can't be, if there can't, something which we might call 'quasation'—just like causation except for the unidirectionality.

All three problems are, on the face of it, just as serious for the counterfactual theory. Consider the problem of effects. If the switch had not been flipped, the light would not have gone on, *and*, if the light had not gone on, the switch would not have been flipped. If the pressure had not been *p*, the barometer would not have read *r* (so, the pressure being *p* caused the barometer to read *r*). What about the reverse? This is what Lewis says about this example:

> If the reading had been higher, would the pressure have been higher? Or would the barometer have been malfunctioning? The second sounds better: a higher reading would have been an incorrect reading.... When [we suppose a higher reading], it is less of a departure from actuality to hold the pressure fixed and sacrifice the accuracy of the barometer, rather than vice versa. It is not hard to see why. The barometer, being more localized and more delicate than the weather, is more vulnerable to slight departures from actuality. (1973/1986a: 169)

There are three problems with what Lewis says here. First, even if Lewis is right about this example, there is no reason why this feature should generalize. Suppose I am recording by means of the sturdiest equipment in the world, the exceedingly volatile behaviour of the molecules in a gas, which are extremely 'vulnerable to small departures from actuality'. If the recording had been different, would the movements have been different? Or would the equipment be malfunctioning? The movements of the particles could *very easily* have been different, whereas it would take a near miracle to make *this* equipment malfunction. We get the wrong result.

Second, Lewis is probably wrong about the facts of his example. As Charles Travis commented to me, *nothing* is more delicate, more vulnerable to slight changes in actuality, than the weather! (There is the well-known 'Butterfly Effect', according to which a butterfly flapping its wings today in Peking can cause storm systems next month in New York. See Gleick, 1987: 8). Whereas barometers can be made as sturdy as you please.

Third, and most seriously, Lewis avoids one troublesome counterfactual only by embracing another. The 'events' (in Lewis's broad sense of the term) of the functioning of the barometer, or the non-functioning of the barometer, are the sort of thing that can cause and be caused. Lewis tells us that if the reading had not been *r*, the barometer would not have been functioning properly. But the reading's being *r* did not *cause* the barometer to be functioning properly! Quite the reverse![4]

If we can't solve the problem of effects, we can't solve the problem of epiphenomena, for there may be a causal chain back to the common cause and forward to the other effect. In any case, counterfactuals sometimes have that pattern. This example I owe to Arif Ahmed: we all ate the same pie for dinner. Fred got ill and died. The question is whether the cause of death was food poisoning from the pie. No, it is argued: suppose the pie was poisonous. Then if Fred had got ill, we would have got ill too (or at least that is quite likely), and we didn't. We test a drug on rats and conclude that it is safe for humans, because, if it were unsafe for humans, it would be unsafe for rats. The counterfactuals are quite in order, yet they have the same structure as arguing from spots to fever, or from barometer to rain.

And of course, there is the problem for the counterfactual theory of the back-up cause waiting in the wings, which would have produced the result had the actual cause failed, the problem of pre-emption. Jane will shoot Mary if and only if John misses, and

[4] I owe this point to Eduardo Flichman (1989).

Jane has perfect aim. However, John doesn't miss. Mary's death does not counterfactually depend on John's shot.

There have of course been many further developments in counterfactual theories of causation since Lewis's pioneering article, but I will leave them aside. It seems to me that the difficulties are symptoms of trying to explain a fundamental feature of the world and our conception of it, in terms of a less fundamental feature. Suppose I was right to say, following Anscombe, that we perceive a lot of causation. However firm our counterfactual beliefs, we do no *perceive* facts about what would have happened if such-and-such had not happened.

3

Consider a class of 'standard' counterfactuals, whose antecedents and consequents are about particular states of affairs holding at particular times, such that the consequent concerns a later time than the antecedent, like 'If you had struck the match it would have lit'—the sort which are most promising for a counterfactual account of causation. How do we assess them? The problem of counterfactuals has always been: what do you hold constant and what do you give up when you make a counterfactual supposition? Suppose the Lewisian picture is approximately right: we consider those *A*-worlds (worlds in which the antecedent, *A*, is true) which are exactly like the actual world until shortly before the antecedent time, and (if the antecedent is actually false) diverge from the actual world at an inconspicuous fork, and obey the laws of the actual world thereafter; and we ask whether the consequent would be true in such worlds. Once we have diverged from the actual world, it is of 'little or no importance' he says (Lewis, 1979/1986a: 48) whether the relevant worlds are approximately similar to the actual world in matters of particular fact.

He needed to say this to avoid what he called the 'future similarity' objection—that his account gives wrong results for any counterfactual of the form: 'if such-and-such had happened, things would have been staggeringly different'; Kit Fine (1975) gave the example, which Lewis discusses, 'If Nixon had pressed the button in 1974, there would have been a nuclear holocaust'. For it is hard to see how, by ordinary standards of similarity, these worlds in which things are staggeringly different will come out the most similar to actuality.

But, pace Lewis, we sometimes do and sometimes don't keep future particular facts constant in assessing counterfactuals, and whether we do depends on whether they are causally independent of the antecedent. I decline to bet on the toss of a coin. It is tossed anyway. It lands heads. If I had bet on heads I would have won. This is so provided that the toss and my betting are causally independent of each other: say the toss takes place in one room, I write Heads or Tails or No Bet on a piece of paper in another room, and there is no causal interaction between these events. If on the other hand, my betting might have caused you to toss the coin a little later, or a little differently, the counterfactual doesn't hold.

Here is an example of this form I have used elsewhere (Edgington, 2004): the car breaks down on my way to the airport. I miss the plane. Later I discover it crashed: if I had caught the plane I would be dead. Suppose that a chance event, not predictable in advance, brought down the plane. At the time of take-off, the plane was not relevantly different, with respect to safety, from any normal plane: there was a small but non-zero chance that some such accident would occur—due to freak weather conditions, or freak electrical or mechanical faults (or combinations thereof), or a freak heart attack or heart attacks on the part of those in control. Suppose also that my absence from the plane had no effect on the causal history of the crash: it's not the case that, e.g. some subtle feature of the distribution of weight contributed to the crash, which might have been different, had I been aboard. In those circumstances, the remark seems correct: we keep constant later features of the actual world, provided that they are causally independent of the antecedent.

Causal independence is a stronger requirement than statistical independence. As a fantasy, imagine that the crash has this genesis: the devil spins a spinner, which has a one-in-a-million chance of landing in the space designated 'crash'. It does land there. There is a crash. If I had caught the plane I would be dead. Now suppose the devil has two identical spinners, and some rule for deciding which to use, which has the consequence that he will spin one if I am on the plane, the other if I am not. Although the chances are initially just the same, in this case, if I had caught the plane, very probably it would not have crashed.

There is a different sort of example, which tempts Lewis to say that future similarity counts for nothing, due to Pavel Tichy (1976). When Fred goes out and it's raining, he always takes his hat. When he goes out and it's not raining, it's a random 50–50 whether he takes his hat. On this occasion, it's raining and he takes his hat. Consider 'If it had not been raining, he would (still) have taken his hat'. A fine-weather world in which he does so resembles the actual world more than a fine-weather world in which he does not. But this, Lewis rightly wants to say, counts for nothing. The counterfactual is not clearly true.

Many examples go the other way. If I had bought these shares, I would be rich. If I had left five minutes earlier, I would have avoided the accident. If I had got up five minutes earlier, the result of the Australian general election would have been just the same.

I pick a coin from a bowl of coins, toss it, and it lands heads. It would be wrong to claim that if I had picked a different coin, it too would have landed heads. But it would be absurd to deny that if Frank in Australia had scratched his nose a minute or two earlier, the coin I picked, which actually landed heads, would still have landed heads. I do not see how we can distinguish these examples without appealing to the notion of causal independence.

Here is another such pair. There is guerrilla warfare in an imaginary country. The guerrilla leader is hiding in a certain village. Government troops have a range of missiles aimed at the village. These devices are indeterministic, and each has a chance of, say,

90% of firing when activated. News having just arrived of the need to deploy troops elsewhere, only one missile is to be set off. The General chooses a missile. It fizzles out. No harm is done.

'We were lucky', says a potential victim later. 'Had the General chosen a different missile, we might well be dead'. His companion, versed in Lewis's earlier work on counterfactuals, demurs. 'We were lucky *it* didn't go off', he says, 'but your relief is misplaced. In the world most like the actual world in which he chose a different missile, it fizzled out too, right? That is to say, if he had chosen a different missile, it too would have fizzled out'. The silliness of this suggests that Lewis should say 'future similarity counts for nothing', as with Tichy's hat.

Version two of the story: two inhabitants of the village are delayed on their way home because they notice a sheep caught in a cactus, and it takes them a while to free it. The scenario is as before, but let me lower the chance each missile has of firing to 25%. This time, the missile does fire. They hear it in the distance. When they get back they meet havoc and destruction. 'If we hadn't noticed the sheep, we would probably be dead now', says one. His companion, versed in Lewis's later work and the reasons for saying 'future similarity counts for nothing', demurs: 'Consider the possible world which deviated from the actual one at the point where we noticed the sheep: the missile (or its counterpart) in that world had only a 25% chance of firing. So if we hadn't noticed the sheep, it's 75% likely that the disaster would not have happened. What a pity we noticed the sheep! If we hadn't, probably all would be well.'

I don't see how Lewis can handle these examples without appealing to the notion of causal independence. Whether Fred wears his hat is not causally independent of the weather. Picking another coin or missile begins another causal process. But the outcome for this coin or missile that was picked is causally independent of someone scratching their nose in Australia, or the antics of the sheep. As Lewis wants to explain causal dependence and independence in terms of counterfactuals, this is a problem for him.

4

Perhaps the most common and most important class of counterfactuals are those that track a causal relation between events. But this is not the only kind of counterfactual judgement we make. Not all indicative conditional judgements track causation, and pretty well any acceptable indicative conditional can 'go counterfactual'—can be transformed into a counterfactual judgement, in a suitable context. Here is an extreme example (borrowed and adapted from Grice, 1989), which concerns a very minimal ground for an indicative. If the shift to the counterfactual is permissible here, it looks as if it is permissible, in a suitable context, for any indicative. There is a treasure hunt. The organizer tells me: 'I'll give you a hint: it's either in the attic or the garden.' Trusting the speaker, I think 'If it's not in the attic it's in the garden'. We are competing in pairs: I go to the attic and tip off my partner to search the garden. I discover the treasure. 'Why did you tell me to go the garden?', she asks. 'Because if it hadn't been in the attic

it would have been in the garden: that's (what I inferred from) what I was told.' That doesn't sound wrong in the context.

Or consider: 'Why did you hold Smith for questioning?' 'Because we knew the crime was committed by either Jones or Smith—if it hadn't been Jones, it would have been Smith'. And if that's allowed, so is the notorious Oswald-Kennedy counterfactual in a suitable context. 'You had already got Oswald. Why did you continue the investigation?' 'We weren't sure it was Oswald; if it hadn't been Oswald, it would have been someone else in the crowd who killed Kennedy'. There's also a nice example of van Fraassen's (1981): the conjuror holds up a penny and claims he got it from the boy's pocket. 'That didn't come from my pocket', says the boy. 'All the coins in my pocket are silver. If that had come from my pocket, it would have been a silver coin.'

Consider any indicative conditional, 'If *A*, *B*', which I accept, on whatever grounds. Suppose I then discover that not *B*, in a way which does not undermine my conditional judgement. It is acceptable for me then to say, 'Not *A*, because if it were the case that *A*, it would be the case that *B*'?

Here is an example from Jonathan Bennett (2003: 343–4) which shows that an indicative conditional, even when one would naturally assume that it involved causation, may be accepted on other grounds: 'If it rains tomorrow the roads will be slippery'. But I don't mean that rain will *make* the roads slippery—the roads are very well constructed and not made slippery by rain. I've just received a leaflet from the council which (a) includes a weather forecast predicting rain; and (b) says they intend to oil the roads tomorrow, warning that this will make the roads slippery. It doesn't look as if it's going to rain, but the council has a first-rate forecaster. However, there is some reason to suspect that the leaflet may be a hoax and not genuine. If it rains, that will be evidence that it is genuine, and hence that they will oil the roads, and hence that the roads will be slippery. Of course, one would mislead by making that conditional remark without warning that the most obvious ground is not the operative one. But that is pragmatics. No conditional that does explicitly use causal language like 'produce' or 'make' or 'result' or 'outcome', forces a causal reading, though it is very often rightly presumed to be asserted on causal grounds. 'If A happens, B will happen, but A won't cause B to happen' is never contradictory.

And if I am right in claiming, in the penultimate paragraph, that any indicative conditional can be transformed into a counterfactual conditional in a suitable context, the same holds for counterfactuals: only if they explicitly have causal content is a causal reading semantically obligatory.

Thus, counterfactuals are too wide a class to hope to capture causation in terms of them. Even for those counterfactuals which are aligned with causal claims, we need to appeal to the notion of causal independence in giving an account of which features of the actual world we hold constant when assessing them (Section 3). Attempts to define causation in terms of counterfactuals are beset by much the same difficulties as

attempts to define causation in terms of regularities or laws (Section 2). This is not surprising because causation is so fundamental a notion as to pervade nearly all our thinking about the world. It is hard to see how we could give an informative analysis of it in terms of something that doesn't presuppose it (Section 1).

References

Anscombe, G.E.M. (1971) 'Causality and Determination'. Reprinted in *Causation and Conditionals*, Ernest Sosa (ed), pp. 63–81. Oxford: Oxford University Press, 1975.

Ayer, A.J. (1972) *Probability and Evidence*. London: Macmillan.

Bennett, Jonathan (2003) *Conditionals*. Oxford: Oxford University Press.

Davidson, Donald (1963) 'Actions, Reasons and Causes', *Journal of Philosophy* 60: 685–700.

—— (1967) 'Causal Relations', *Journal of Philosophy* 64: 691 –703.

Edgington, Dorothy (2004) 'Counterfactuals and the Benefit of Hindsight'. In: *Cause and Chance: Causation in an Indeterministic World*, P. Dowe & P. Noordhof (eds), pp. 12–27. London: Routledge.

Evans, Gareth (1985) 'Things Without the Mind'. In his *Collected Papers*, pp. 249–90. Oxford: Oxford University Press.

Fine, K. (1975) 'Review of Lewis's *Counterfactuals*', *Mind* 84: 451–8.

Flichman, Eduardo H. (1989) 'The Causalist Program: Rational or Irrational Persistence?', *Crítica* 21(62): 29–53.

Gleick, James (1987) *Chaos*. London: Penguin Books.

Goldman, Alvin (1967) 'A Causal Theory of Knowing', *Journal of Philosophy* 64: 357–72.

Goodman, Nelson (1946) *Fact, Fiction and Forecast*. Indianapolis: Bobbs-Merrill.

Grice, H.P. (1961) 'The Causal Theory of Perception', *Proceedings of the Aristotelian Society*, Supplementary Volume 35: 121–52.

—— (1989) 'Logic and Conversation'. In his *Studies in the Way of Words*, pp. 22–40. Cambridge, MA: Harvard University Press.

Hempel, C.G. (1965) *Aspects of Scientific Explanation*. New York: Macmillan.

Hume, David (1739) *A Treatise of Human Nature*. London

Hume, David (1748) *Enquiry Concerning Human Understanding*. London.

Kripke, Saul (1980) *Naming and Necessity*. Oxford: Basil Blackwell.

Lewis, David (1973/1986a) 'Causation', *Journal of Philosophy* 70: 556–67, reprinted with Postcripts in Lewis (1986a), pp. 159–213. Page references are to this volume.

—— (1979/1986a) 'Counterfactual Dependence and Time's Arrow', *Nous* 13: 455–76, reprinted in Lewis (1986a), pp. 32–66. Page references are to this volume.

—— (1986a) *Philosophical Papers*, Volume II. Oxford: Oxford University Press.

—— (1986b) 'Causal Explanation', in Lewis (1986a), pp. 214–40.

—— (1986c) 'Events', in Lewis (1986a), pp. 241–69.

Peacocke, Christopher (1985) *Thoughts: An Essay on Content*. Oxford: Basil Blackwell.

Shoemaker, Sydney (1980) 'Causality and Properties'. In: *Time and Cause*, P. van Inwagen (ed.), pp. 109–35. Dordrecht: Reidel.

Stalnaker, Robert (1984) *Inquiry*. Cambridge, MA: MIT Press.

Strawson, P.F. (1985) 'Causation and Explanation'. In: *Essays on Davidson, Action and Events*, B. Vermazen & M. Hintikka (eds), pp. 115–36. Oxford: Oxford University Press.

Tichy, Pavel (1976) 'A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals', *Philosophical Studies* 29: 271–3.

van Fraassen, Bas (1981) 'Essences and Laws of Nature'. In: *Reduction, Time and Reality*, R. Healey (ed.), pp. 189–200. Cambridge: Cambridge University Press.

12

Suppositions, Conditionals, and Causal Claims

Aidan Feeney and Simon J. Handley

1 Introduction

Over several years we have been interested in how people understand and test conditional assertions (for example see Feeney & Handley, 2000; Handley & Feeney, 2004, 2007). One of us has been involved in the development of the suppositional theory of conditionals (see Evans, Over, & Handley, 2005), and in this chapter we will describe how the methods that have been used to test this theory can be turned to examining the processes involved when people think about the causal relation that holds between events in the world. The suppositional theory holds that to understand a conditional one temporarily adds the antecedent to one's store of beliefs and evaluates the consequences. At the core of the theory is the idea that understanding conditionals requires us to engage in mental simulations (see Kahneman & Tversky, 1982; Mandel, this volume). For example, to understand a conditional such as 'If petrol goes to £2 a litre than there will be fewer cars on the road', the theory holds that people imagine a world in which petrol costs £2 per litre in order to see whether this change to the actual world results in fewer cars on the road. This process of mental simulation means that when people are asked how likely it is that the conditional is true, the answers they give should equal their perception of the probability that there will be fewer cars on the road given that petrol goes to £2 per litre.

Although the account claims that people's beliefs about the truth of conditionals depends in large part on the mental simulation of a world in which the antecedent holds, there is some evidence that people also consider possibilities in which the antecedent does not hold when judging the truth of causal conditionals (see Over et al. 2007). In two experiments we will focus on the extent to which simulation of a world in which the antecedent does not hold is related to people's judgements about the truth of causal and concessive (*even if*) conditionals, and to their judgements of causal strength. To anticipate our results, we will show that beliefs based upon 'undoing' the antecedent event are positively related to judgements about concessive

conditionals, and negatively related to judgements of causal strength. However, such beliefs are relatively weakly related to judgements about the truth of causal conditional claims. Our focus of interest is the conditional, but we hope that our results will be of interest to other researchers working in the area of causality where consideration of cause-absent cases is central to a variety of accounts of how relationships between putative causes and their effects are assessed (see Hume, 1748; Lewis, 1973).

Much work in the psychological literature on the relationship between counterfactual and causal thinking has tended to focus on claims about events where the outcomes are known (for a review see Mandel, this volume). As we will shortly see, our work has primarily (but not exclusively) been about causal conditionals concerning future events. For this reason, the possibilities which participants might consider when evaluating such conditionals are more akin to future hypotheticals than to what are commonly thought of as counterfactuals in the psychological literature. Although there appear to be important developmental effects caused by the difference between counterfactual conditionals whose antecedents are false and future hypotheticals (see Perner & Rafetseder, this volume), alongside Woodward (this volume) we argue that future hypotheticals can be closely tied to an understanding of causal judgements. Our focus on future hypotheticals complements existing work on the role of counterfactual thinking in cause ascription for events that have already happened.

By way of introduction we will give an account of the suppositional theory, the methods that have been used to test the theory, and the results of experiments using those methods to examine different conditional forms. One aim of the experiments was to study concessive conditionals (*even if* conditionals) under the suppositional framework. Concessive conditionals are interesting in this context because intuitively they seem to deny rather than to assert causal claims (see Byrne, 2005). Once we have surveyed the literature on the suppositional theory we will consider this type of conditional form in detail and how it might inform our understanding of how people think about causal relations.

2 The Suppositional Account of Conditionals

The framework we will use to consider conditionals in this chapter is known as the suppositional account (see Evans & Over, 2004; Evans, Over, & Handley, 2005; Handley et al. 2006). This account draws upon Ramsey's idea (1931) that ordinary conditionals are evaluated by adding 'p' to one's stock of beliefs and evaluating 'q' in that context. According to the Ramsey test conditionals are believable to the extent that q is probable given p. This view forms the basis of the suppositional account, which claims that when people understand conditionals they engage in a process of mental simulation, which involves simulating the hypothetical world in which p holds and evaluating one's belief in q in the light of this simulation (see, for example, Handley, Evans, & Thompson, 2006).

One of the key predictions that arises from this account is that people should evaluate the believability of a conditional statement as a function of the probability of the consequent in the light of the antecedent [P(q/p)]. The 'conditional probability' hypothesis has been tested in a number of recent studies. In these studies the typical approach involves asking participants to provide a probability judgement concerning the truth of a given conditional, for example:

(1) If less violence is shown on television, then the amount of violent crime will reduce.

Later participants are presented with a 'probabilistic truth table task' (PTT), where they are asked to assign percentage probabilities to each of the four truth table cases associated with the conditional statement. This truth table method is likely to be new to readers unfamiliar with the recent literature on conditionals, but it bears some resemblance to methods used in the literature on contingency learning. The key difference is that the truth table task is used to assess people's beliefs about future hypothetical events whereas in a contingency learning experiment people are presented, either on a case-by-case basis, or in summary form, with information about the rates of occurrence of the four contingency table cases (i.e. cause and effect present; cause present, effect absent; cause absent, effect present; cause and effect absent) relevant to the assessment of a contingency. Thus, in a truth table task, participants are asked to assign probabilities to each of the four conjunctions corresponding to the truth table cases for the conditional in (1), summing to 100%:

(TT) Less violence is shown on television, the amount of violent crime reduces ___

(TF) Less violence is shown on television, the amount of violent crime does not reduce ___

(FT) No less violence is shown on television, the amount of violent crime reduces ___

(FF) No less violence is shown on television, the amount of violent crime does not reduce ___

100%

Note that because the conditionals used in these studies are future hypotheticals, participants are asked to express their beliefs about the probability that each conjunction of events will occur. In a contingency learning experiment participants are told about the rate at which the four relevant conjunctions of cause and effect have occurred.

Based upon participant's responses in the truth table task, it is possible to calculate a range of conditional probabilities, drawing upon an individual's beliefs about the likelihood of various outcomes pertaining. For example, conditional probability, P(q/p) (the probability that violent crime will reduce given that less violence is shown on television) is a function of the probability assigned to the TT case weighted

by the total probability assigned to TT and TF cases (TT/TT+TF). It is also possible to evaluate people's beliefs about the probability that the consequent will occur in the absence of the antecedent (P(q/not-p)) which is a function of the probability assigned to FT cases in proportion to FT and FF cases combined (FT/FT+FF). Despite the emphasis here on probabilistic calculations, it is important to remember that although we infer P(q/p) and P(q/not-p) from the truth table data, the suppositional account is intended to explain the mental representations underlying people's interpretation of conditional statements rather than merely working as an account of the probabilistic calculations underlying people's probability of truth judgements for conditionals. In other words, although the truth table method works as a test of the account, the account does not hold that people have a probabilistic truth table in their head.

Research using this method has provided strong support for the conditional probability hypothesis and the suppositional account of conditionals. Conditional probability has been shown to be strongly correlated with judgements of the probability of abstract conditionals and ordinary conditionals (Evans, Handley, & Over, 2003; Over et al. 2007). For example, Over et al. (2007) reported a study testing whether the conditional probability hypothesis also holds for counterfactual conditionals. In that experiment, participants were shown causal counterfactual conditional statements such as (2):

(2) If New York had not been attacked by terrorists in 2001, then the US would not have attacked Iraq.

Participants were asked to judge the probability that the conditionals were true. The truth table task was modified for this experiment so that participants were asked to estimate the probability of each of the truth table cases from the perspective of a point in time before the events referred to had occurred. The results of this study, once again, contained strong support for the conditional probability hypothesis. That is, P(q/p) was the strongest predictor of the judged probability of a set of 32 counterfactual conditionals. This result is consistent with the claim that in judging the probability of counterfactual conditionals people imagine themselves at a point in time before the events described became counterfactual and simply apply a Ramsey test via a process of mental simulation. This finding suggests, contrary to other claims in the literature, that counterfactuals are understood in a similar way to conditionals that refer to future hypothetical events.

Typically, recent studies have used causal conditionals of the kind shown in (1) and (2) above, where there is a hypothesized causal relationship that holds between the antecedent and consequent events. Now, if experimental participants understand these conditionals as asserting a causal relation, then other probabilities may become important. In particular, because several approaches to causation hold that P(q/not-p) is related to the strength of the causal relationship that holds between p and q, if people are assessing the strength of the causal relationship when interpreting the conditional, then we might expect P(q/not-p) to be negatively correlated with their judgements

about the probability that the conditional is true. As P(q/not-p) and P(q/p) are statistically independent (entirely different truth table cases go into their calculation), it is relatively easy to test this hypothesis. There is mixed support for the proposed negative relationship, with some evidence for a weak negative relationship between P(q/not-p) and probability of conditional judgements. For example, in Experiments 1 and 2 of Over et al. (2007), P(q/not-p) was not significantly correlated with people's judgements of the probability that the causal conditionals were true. Furthermore, the correlation was positive rather than negative in both experiments. However, Experiment 3 of that paper concerned counterfactuals and produced a negative and significant correlation between P(q/not-p) and judgements about the probability of the conditional, and a significant and negative correlation ($r = -.31$) was reported by Evans et al. (2007a).

One possible reason for these mixed results is that in all of the studies described, the correlations were calculated over a relatively small sample of 32 conditionals. One of the aims of the experiments we will describe below was to increase the size of the sample of conditionals. Under such circumstances, if there are reliable effects of P(q/not-p) then we should have sufficient power to detect them. Most studies to date have examined participant's beliefs about the truth of causal conditional statements. In the first experiment presented here we asked for an evaluation of causal strength as a more direct measure of beliefs about causal relations.

Certain conditional expressions, rather than expressing a causal relationship, expressly deny that a causal relationship exists. Even-if or concessive causal conditionals are used when there is reason to question that an expected relationship holds. Consequently one would expect even-if assertions to show opposite patterns of relationship to P(q/not-p) than causal judgements. One of the primary aims of the experimental work that we will shortly present was to test this intuition. However, before describing the experimental work, it is necessary to introduce some of the relevant philosophical and linguistic analyses of even-if conditionals.

3 Concessive or 'Even If' Conditionals

Even-if conditionals invite the listener to suppose that a relation that one might have expected to hold in the world, in fact does not. Consider, for example, the following assertion:

(3) Even if the US had ratified the Kyoto protocol, carbon emissions would have increased.

Such assertions are referred to as semifactuals, because they describe a world in which the consequent event is assumed to have pertained (i.e. carbon emissions have increased), but the antecedent action was not carried out (the US did not ratify the Kyoto agreement). Such conditional assertions are interesting from the perspective of

causal reasoning because they are often used to deny a causal link between an antecedent and consequent event (in this case the link between ratifying Kyoto and carbon emissions). Even-if assertions operate equally well when referring to future events, for example:

(4) Even-if the Conservatives win the election there will be an increase in taxes,

where the outcome of the election is unknown (at the time of writing), but the implication is that taxes will increase irrespective of what outcome pertains, once again denying a causal link between election outcome and tax increases.

A number of authors have emphasized that in order to understand how people represent and reason from even-if, one must first consider the function of *even* in everyday natural language. Consider, for example, the assertion in (5):

(5) Even Tony distrusts George.

Several philosophers (Jackson, 1987; Sanford, 1989) have suggested that *even* serves to deny an available presupposition; for example, that we might expect Tony to trust George. It serves to pick out an extreme position and calls up a range of contextually determined alternatives that are less surprising; for example, that Gordon distrusts George, Barrack distrusts George, or Jacques distrusts George. In so doing the utterance invites the listener to infer that George is a man not to be trusted.

Whilst there exist exceptions, most authors have treated *even-if*, not as a distinct logical connective, but as a construction that consists of a combination of the focusing particle *even* and the *if* of a conditional. The analyses presented above can readily be extended to *even-if*. To illustrate consider the statement in (4) above. The assertion denies a general presupposition that the Conservatives are a tax reducing party; that is, it denies an association of the following kind:

(6) If the Conservatives win the election then taxes will be reduced.

As we have seen *even* picks out an extreme possibility on a scale of related statements that are more probable in a given context. This provides a series of effects related to unexpectedness or surprise. Declerck & Reed (2001) have argued that *even-if* similarly induces a sense of unexpectedness. This sense of unexpectedness relates to the conditional as a whole and leads to what they label an expectation understanding, i.e. one might expect p to preclude q (a Conservative victory to preclude tax rises), and a non-preclusive understanding, p does not (in fact) preclude q (in fact a Conservative victory will not preclude tax rises). The focal conditional lies at the extreme point of a scale of unexpectedness, consequently calling to mind a range of conditionals that are more likely and less surprising, such as:

(7) If the Liberal Democrats win the election then there will be tax rises,
(8) If the Labour party wins the election then there will be tax rises.

These less surprising conditionals license the inference that the conditional relationship holds for all other values of the same scale and hence for a series of antecedents (Konig, 1986). According to Jackson (1987; see also Declerck & Reed, 2001) the range often consists of the conditional with the antecedent negated:

(9) If the Conservatives don't win the election then taxes will rise.

The combination of (9) with the assertion in (4) leads to the inference that taxes will rise whether or not the conservatives win the next election. In formal terms this inference is what logicians would term *constructive dilemma* and it corresponds to the intuition that many *even-if* assertions appear to entail their consequent.

4 A Psychological Account of 'Even If'

A key question in this paper concerns the component probabilities, and the associated mental simulations, which predict judgements about the probability of even-if conditionals. Before outlining our predictions we will briefly present our psychological account of even-if, which draws heavily on the linguistic accounts described earlier combined with the suppositional account of conditionals, described above (for an alternative account see Byrne, 2005; Moreno-Rios, Garcia-Madruga, & Byrne, 2008). Consider, to start with, the following causal conditional:

(10) If the Conservatives win the election then there will not be tax rises.

Of course one might not believe (given a bleak economic outlook) that any sort of relationship holds between the election outcome and tax reductions. These conditions create a perfect opportunity for asserting an even-if conditional of the kind shown in (11):

(11) Even-if the Conservatives win the election there will be an increase in taxes.

As argued above, even-if calls up a range of alternative conditionals on a probability scale that are less surprising or unexpected such as the conditionals in (7) and (8). In probabilistic terms, the probability of the consequent (an increase in taxes) is greater given this range of alternative antecedents (such as a Labour or Liberal victory), compared to the probability of the consequent given the antecedent in (11). Thus, the conditions required to support an even-if assertion will be the opposite of those required to support the assertion of a causal conditional. Specifically an even-if assertion will be assertable under conditions where P(q/not-p) is high. These are the very conditions under which a causal relationship is called into question, thus supporting the general intuition that even-if, not only serves to deny an existing presupposition, but is also used to cast doubt on a causal relationship between an antecedent (election outcome) and a consequent event (tax outcome).

In the two experiments that follow we evaluate the extent to which judgements concerning causal strength, the probability of even-if statements, and the probability of

causal conditionals are influenced by various conditional probabilities. In Experiment 1 we compare causal strength judgements with judgements concerning the truth of even-if assertions. We predict that the conditional probability hypothesis will be supported in both cases, and that P(q/not-p) will be negatively associated with judged causal strength, but positively related to judgements about the probability of even-if conditionals.

Experiment 1

Sixty-four undergraduate student volunteers from the University of Plymouth took part in Experiment 1 in return for participant payment. There were twenty-four men and forty women whose ages ranged from 19 to 45 years of age. We employed a within-participants design in which all participants completed the probabilistic truth table task (PTT) and one of two judgement tasks. One of these tasks required participants to rate the probability that a series of 'even-if' statements was true. This task was modelled on the conditional judgement task as used by Over et al. (2007). Participants who completed the other task were asked to estimate the strength of the causal link between the antecedent and the consequent of the base conditionals from which the even-if sentences were constructed. Each task was presented in a booklet, with each booklet containing items derived from the same base set of fory-eight conditional statements (see Table 12.1 for examples), and participants were randomly allocated to one of the two conditions.

The even-if conditionals were constructed from a set of base conditional statements by adding a negation to the consequent clause. An example base, and even-if, conditional, along with the corresponding causal statement, is shown below:

Base Conditional: If car ownership increases, traffic congestion will get worse.
Even If Conditional: Even if car ownership increases, traffic congestion will not get worse.
Causal Statement: An increase in car ownership will cause traffic congestion to get worse.

Table 12.1 Example sentences used in Experiments 1 and 2

1. If oil prices continue to rise, then UK petrol prices will rise.
2. If car ownership increases, then traffic congestion will get worse.
3. If high-risk prisoners are released early, then the crime rate will increase.
4. If Adidas gets more superstars to wear their football boots, then sales will increase.
5. If children are paid to go to school, then attendance levels will increase.
6. If parenting is taught in schools, juvenile crime rates will increase.
7. If footballers' wages increase, then more goals will be scored during the football season.
8. If compulsory PE lessons are introduced until the age of 18, then child levels of obesity will rise.
9. If EU quarantine laws are strengthened, rabies will spread to the UK.
10. If the legal age of driving is lowered, then the number of accidents on the road will fall.

The forty-eight statements in the conditional probability and causal strength judgements tasks were presented in a random order for each participant. Participants were told that they would be presented with a list of statements relating to events that may occur in the next 10 years. In the conditional judgement task they were asked to provide an estimate of the probability that each sentence is TRUE, on a percentage scale. Participants who received the causal statements were asked to indicate how strongly they judged the causal link between the two events described. They indicated their judgement on a five-point scale with the end points labelled 'weak causal link' and 'strong causal link', respectively.

In the probabilistic truth table task (PTT) participants are required to assign percentage probabilities to each of the four conjunctive truth table cases corresponding to each conditional. For example, the truth table cases presented to participants corresponding to the conditionals above were:

(TT) Car ownership increases and traffic congestion gets worse ___
(TF) Car ownership increases and traffic congestion doesn't get worse ___
(FT) Car ownership doesn't increase and traffic congestion gets worse ___
(FF) Car ownership doesn't increase and traffic congestion doesn't get worse ___
100%

Participants were asked to judge the probability that the specified events would occur in the next 10 years. They were instructed to assign a probability to each of the truth table cases with a requirement that the probabilities summed to 100%. Forty-eight sets of truth table cases were presented corresponding to each of the forty-eight conditional statements. These were presented in a random order for each participant. Half of the participants received the conditional judgement task followed by the probabilistic truth table task, with the remaining participants receiving the tasks in the reverse order.

Findings. The even-if assertions were constructed from the base conditionals by adding a negation to the consequent clause as indicated above. For clarity we calculated the conditional probabilities relative to the Antecedent (A) and Consequent (C) clauses, regardless of negation. For example, given the even-if assertion in Table 12.2, 'Even-if car ownership increases traffic congestion will not get worse', P(C/A) represents the probability that traffic congestion will not get worse given an increase in car ownership, whereas for the conditional, 'If car ownership increases then traffic congestion will get worse', P (C/A) represents the probability that traffic congestion will get worse given an increase in car ownership. This approach allows us to directly examine the extent to which logically equivalent probabilities predict probability and causal strength judgements in different ways for the different tasks.

The analyses of interest here concern probabilities calculated on the PTT task and their relationship to judged probabilities assigned to the conditional sentences, and ratings of the strength of the causal link between antecedent and consequent. The

Table 12.2 Example statements with associated Antecedent (A) and Consequent (C), for each of the three tasks used in Experiments 1 and 2

Task	*Example statement*	*Antecedent (A)*	*Consequent (C)*
Causal judgement	An increase in car ownership will cause traffic congestion to get worse	Car ownership increases	Traffic congestion gets worse
Even if	Even if car ownership increases, traffic congestion will not get worse	Car ownership increases	Traffic congestion does not get worse
Causal conditional	If car ownership increases, traffic congestion will get worse	Car ownership increases	Traffic congestion gets worse

analysis we report collapses over participants, relating these probabilities as measured across problem materials, in line with previous research using this method (see Over et al. 2007). This was achieved by first computing the mean for all participants for each sentence and then calculating correlations and regressions on these mean scores. Table 12.3 shows the raw correlations between each of two relevant probabilities (P(C/A), and P(C/not-A)) and both causal strength ratings and probability judgements for even-if sentences.

Looking first at the raw correlations corresponding to the conditional probability hypothesis outlined earlier, it is clear that the conditional probability of the consequent given the antecedent is significantly associated with causal judgements and with probability judgements for even-if sentences. However, even-if is not predicted as well by P(C/A) as are causal judgements, and whereas P(C/not-A) is significantly, and negatively, associated with causal judgements, the association with even if judgements is non-significant.[1] Recall that one proposed function of even-if is to deny a causal link between the antecedent and consequent clauses. On this basis one might expect P(C/not-A) to be positively correlated with the probability of even-if, given that a weak causal relation provides stronger conditions for asserting a concessive conditional.

Table 12.3 Correlations between probability judgements and derived probabilities calculated over sentences in Experiment 1

	Causal Strength	Even-if
Predictors		
P(C/A)	.96**	.76**
P(C/not-A)	-.39**	.08

[1] One difficulty in making a qualitative comparison between these correlations is that causal strength was measured on a 5-point scale whereas probability of truth was measured on a 100-point scale. Therefore, it is important to note that participants' responses on both scales were normally distributed and that variance on one was not more restricted than variance on the other.

Table 12.4 Regression models for even-if and causal strength judgements from Experiment 1 based upon two independent predictors

	Beta weights for predictors		F Value	R^2 adjusted
	P(C/A)	P(C/not-A)		
Causal	.92**	-.15**	F (3,44) = 381.8, p<.001	.94
Even-if	.84**	.29**	F (2,45) = 43.6, p<.001	.64

Note: *p < .05, ** <.01

The correlational data only weakly support this analysis. In order to provide a more robust test of our proposals we used the approach adopted in Over et al. (2007), performing multiple linear regression analyses with judged probability as the dependent variable with two statistically independent predictors: P(C/A) and P(C/not-A).

To the extent that judgements of the strength of the causal link between the antecedent and the consequent of a causal conditional draws on the same processes as judging the probability of a causal conditional, the conditional probability hypothesis predicts a strong effect of P(C/A) on judgements of causal strength. Whereas P(C/A) should be a positive predictor of causal judgements, P(C/not-A) should be a negative predictor. For 'even-if' conditionals on the other hand, P(C/not-A) should be a positive predictor as a high value of P(q/not-p) provides the conditions for making an assertion of this kind. Table 12.4 shows the beta weights and regression models for each type of judgement for both predictors.

The regression models for causal and even-if judgements are both highly significant explaining 94% and 64% of the variance respectively. As predicted, conditional probability, P(C/A), was a significant positive predictor of causal strength judgements, whilst P(C/not-A) was a negative predictor. It is worth noting that the beta weight for P(C/A) is much larger than the beta weight for P(C/not-A), indicating that conditional probability is overweighted in an evaluation of causal strength relative to a normative model such as the delta p rule (see Cheng, 1997). Turning now to ratings of even-if, as with causal judgements, P(C/A) had a large and reliable beta weight. In line with the account outlined earlier, this was accompanied by a smaller, but highly reliable positive beta weight for P(C/not-A), demonstrating that the ratings of even-if increase as the probability of the consequent occurring in the absence of the antecedent increases. This quite clearly indicates that even-if cues people to think about alternative causes, thus undermining the link between the antecedent event and the stated consequent.

The data thus far provide good evidence to suggest that causal and even-if judgements differ in the way in which they cue people to think about alternative causes. In contrast to causal judgements, even-if assertions are rated as more probable the greater the probability that the consequent will occur in the absence of the antecedent. As predicted by a number of accounts of causality, causal strength is negatively related to the judged probability that the consequent will occur in the absence of the antecedent.

We have found that P(C/A) is a positive predictor of judgements about even-if and we argue that this finding relates to the important function served by even-if in denying a presupposed relationship. Declerck & Reed (2001) have argued that even-if denies the presupposition that p would normally prevent q; that is p would normally lead to not-q. This presupposition clearly relates to an understanding of the relation that would normally hold between the antecedent and consequent. Therefore, one might expect the acceptability of an even-if assertion to be related to the ease by which the presupposition can be activated. It is of course well known that the ease of activation of a denied presupposition depends upon the extent to which the denial is a plausible one (see, for example, Wason, 1965). The data suggest that conditional probability is important in determining the acceptability of even-if assertions, but this may only be up to a certain point, because the higher the conditional probability, the less available a relevant presupposition will be; it will no longer be a case of a 'plausible denial'. To illustrate this, consider the assertion in (12):

(12) Even if Brazil score five goals they will win the game.

There is something odd about even-if assertions of this kind where the conditional probability is particularly high because the presupposition, 'If Brazil score five goals then they will lose the game' is inconsistent with our beliefs about football matches in general and Brazilian football teams in particular. Thus, this presupposition is unlikely to be highly available to us when we process the even-if assertion. This example suggests that there may not be a simple linear relationship between conditional probability and the probability of even-if sentences. To test this hypothesis we carried out further analysis of the relationship between judgements for even if and P(C/A). Figure 12.1 presents a scatterplot showing the relationship between these two measures calculated across all 48 even-if sentences. Our analysis suggests that even-if sentences will be most acceptable where P(C/A) is neither at the extreme low- or high-end of the scale. This in turn suggest that the relationship between P(C/A) and sentence ratings will be curvilinear. Table 12.5 shows a comparison between a linear and quadratic curve fitting to the data. Although the linear model was a good fit, it was significantly outperformed by the quadratic model.

Experiment 1 employed the probabilistic truth table method to enable us to identify the underlying beliefs that relate to causal judgements and to the judged probability of even-if conditionals. The findings are relatively clear; the acceptability of even-if is linked to our understanding of causality. As we have argued, the weaker the causal relationship between p and q, the more acceptable even-if should be. However, as a conditional, even-if's assertability conditions require that P(q/p) is also at an acceptable level. For a causal relation to be weak, P(q/not-p) must also be high. Our analysis confirms that both of these conditional probabilities are important in predicting the acceptability of an even-if assertion linking them squarely with an understanding of causality. In addition, in Experiment 1 there was some evidence that P(q/not-p) is related to the strength of the causal relationship that is perceived to exist between p and q. This finding is consistent with a number of theories, including counterfactual theories (see Lewis, 1973), of causation.

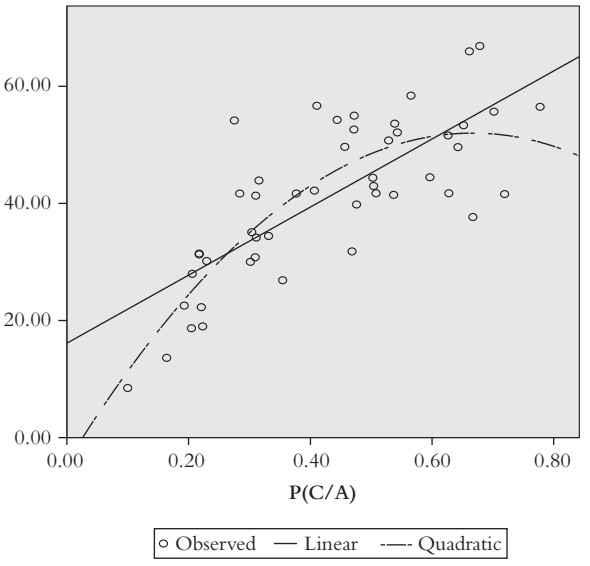


Figure 12.1 The relationship between P (C/A) and sentence ratings for even-if in Experiment 1.

Table 12.5 Comparison of linear and quadratic curve fitting for even-if against P(C/A) in Experiment 1

	Beta weights for predictors		F Value	R^2 adjusted
	P(C/A)	$P(C/A)^2$		
Linear	.76**	–	F (1,46) = 63.5, p<.001	.57
Quadratic	.82**	.28**	F (2,45) = 42.7, p<.001	.64
R^2 change test: F(1,45)=9.78, p<.001				

Note: *p < .05, ** <.01

In Experiment 2 we asked participants to judge the probability that a series of future hypothetical conditionals are true. The aim of Experiment 2 was to compare judgements concerning the acceptability of even-if assertions and if assertions, evaluating the extent to which P(q/not-p) predicted differentially in each case. Experiment 2 also allowed us to attempt to replicate the striking curvilinear relationship we observed in Experiment 1 between P(C/A) and judgements about even-if.

Experiment 2

Seventy undergraduate student volunteers from the University of Durham participated in Experiment 2 in return for payment. There were sixteen men and fifty-four women aged from 18 to 49 years. We employed a within-participants design in which all

participants completed the probabilistic truth table task (PTT) and a conditional judgement task as used by Over et al. (2007). Each task was presented in a booklet, with each booklet containing items derived from the same base set of forty-eight conditional statements (see Table 12.1). Participants were randomly allocated to one of two conditions, providing ratings for conditionals containing the connectives 'if then' (IT), or 'even-if' (EI). As was the case in Experiment 1, the even-if versions were constructed from the base conditional statements by adding a negation to the consequent clause. The forty-eight statements in each of the conditional judgement tasks were presented in a random order for each participant. In the probabilistic truth table task participants are required to assign percentage probabilities to each of the four conjunctive truth table cases corresponding to each conditional. The forty-eight sets of truth table tasks were presented in a random order for each participant. Half of the participants received the conditional judgement task followed by the probabilistic truth table task, with the remaining participants receiving the tasks in the reverse order.

Findings. Table 12.6 shows the raw correlations between the inferred probability values and people's judgements about even-if and if conditionals. Note that the pattern of correlations for even-if is very similar to those reported in Experiment 1 and the correlations between the predictors and judgements about causal conditionals are very similar to those reported for causal judgements in Experiment 1. Conditional probability is strongly positively correlated with both judgements. It is also worth noting that P(C/not-A) is a negative predictor for causal conditional judgements as expected, but as shown in Experiment 1, is uncorrelated with even-if judgements.

As in Experiment 1, the pattern of correlations provides mixed evidence to suggest that causal conditional and even-if judgements are related to the calculated conditional probabilities in different ways. Once again we adopted a regression solution by examining the extent to which our two independent predictors, P(C/A) and P(C/not-A), predicted judged probability of each sentence.

The conditional probability hypothesis predicts a strong effect of P(C/A), and if causal strength is a factor then P(C/A) should be a positive predictor and P(C/not-A) should be a negative predictor for 'if then' conditionals (which may, as a part of their meaning, assert a causal relation). On the other hand, P(C/not-A) should be a positive predictor for 'even-if'. Table 12.7 shows the beta weights and regression models for each type of sentence for both predictors.

Table 12.6 Correlations between probability judgements and derived probabilities in Experiment 2 calculated over sentences

	If then	Even-if
Predictors		
P(C/A)	.96**	.83**
P(C/not-A)	-.45**	.06

Table 12.7 Regression models for even-if and if-then in Experiment 2 based upon two independent predictors

	Beta weights for predictors		F Value	R^2 adjusted
	P(C/A)	P(C/not-A)		
If-then	.92**	-.09†	F (2,45) = 249.9, p<.001	.91
Even-if	.96**	.32**	F (2,45) = 81.56, p<.001	.77

Note: †p < .1, *p < .05, ** <.01

Table 12.8 Comparison of linear and quadratic curve fitting for even-if against P(C/A) in Experiment 2

	Beta weights for predictors		F Value	R^2 adjusted
	P(C/A)	$P(C/A)^2$		
Linear	.83**	–	F (1,46) = 105, p<.001	.69
Quadratic	.94**	.33**	F (2,45) = 86.8, p<.001	.78
R^2 change test: F(1,45) = 21.41, p<.001				

Note: *p < .05, ** <.01

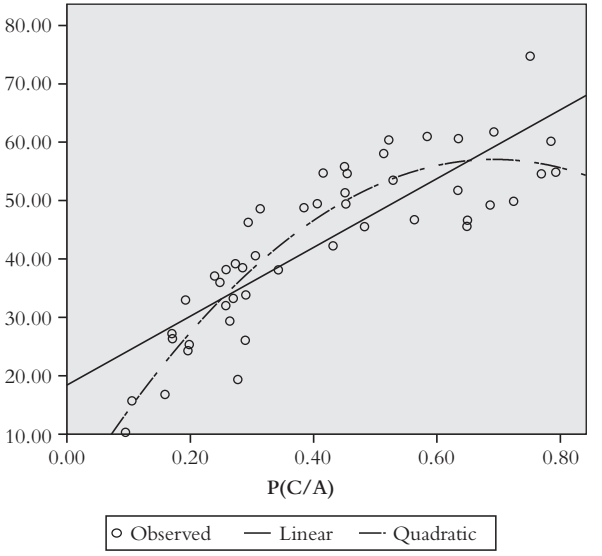


Figure 12.2 The relationship between P(C/A) and sentence ratings for even-if in Experiment 2.

In line with earlier findings, this analysis revealed strong support for the conditional probability hypothesis for causal conditionals, with a highly significant regression model and P(C/A) having by far the highest beta weight. However, there was weak support for the causal hypothesis, with P(C/not-A) eliciting only a marginally significant beta weight. Turning now to the ratings of the even-if sentences, as Table 12.7 shows, the regression model was highly significant. As with causal conditionals, P(C/A) had a large and reliable beta weight. In line with the findings of Experiment 1, this was accompanied by a smaller, but highly reliable positive beta weight for P(C/not-A).

Finally we examined the relationship between P(C/A) and ratings of even-if statements, in order to determine whether the curvilinear relationship reported in Experiment 1 was replicated in this data set. Table 12.8 shows a comparison of the linear and quadratic curve fitting and once again demonstrates a significant advantage of the curvilinear over the linear model, a pattern illustrated by the scatterplot in Figure 12.2.

5 Back to Causality

The two experiments we have described here contain a number of interesting findings. In Experiment 1 we examined people's beliefs about the strength of the causal relationship between the antecedent and consequent events together with beliefs about the truth of even-if assertions concerning the same relations. The findings show that whereas conditional probability is a positive predictor of both types of judgement, the probability of the consequent given the absence of the antecedent is a negative predictor of judgements of causal strength and a positive predictor of probability judgements for concessive conditionals. Two things should be noted here. First, P(C/A) was a much stronger predictor of causal judgements than was P(C/not-A). Second, P(C/A) best predicted probability judgements for concessives when it was neither so low as to make the concessive implausible nor so high as to make it infelicitous.

In Experiment 2 we compared probability judgements for concessive and causal conditionals. The results for even-if were consistent with Experiment 1. Thus, P(C/A) and P(C/not-A) were positive and significant predictors of participants' judgements about the probability that 'even if A then C' was true, and there was a curvilinear relationship between P(C/A) and P(even if A then C). For causal conditionals, on the other hand, the results strongly supported the conditional probability hypothesis, with P(C/A) being very strongly predictive of P(if A then C). However, there was only weak evidence to suggest that when considering the probability that a causal conditional statement is true people consider hypothetical possibilities in which the antecedent is false. Thus, whilst beliefs based upon 'undoing' the antecedent event are positively related to judgements about concessive conditionals, and negatively related to judgements of causal strength, they appear relatively weakly related to judgements

about the truth of causal conditional claims. Previous research using this paradigm has produced inconsistent evidence to suggest that P(C/not-A) is related to beliefs in causal conditionals. Our findings, which are based upon a substantially greater set of conditional sentences, suggest that if the effect is present it is weak.

These results speak to the relationship between counterfactuals and causal claims in a number of ways. Most obviously they suggest that when evaluating the truth of causal claims expressed as indicative conditionals, people rarely consider the false antecedent possibility identified as important by theories of causation. That is, people don't appear to think about the likelihood of the effect in the absence of the cause. However, when participants are explicitly asked to consider the strength of the causal relationship between the antecedent and consequent of such causal claims people do think about whether the effect is likely to occur in the absence of the putative cause. Of course, as we pointed out earlier, considerations of false antecedent cases do not predict people's causal strength judgements as strongly as do considerations of true antecedent cases. This could be because people underweight the cause-absent possibilities, which would be consistent with their tendency to underweight certain information when judging contingency (see Anderson, 1990; Shanks, 1995, Stanovich & West, 1998). Specifically, people tend to pay more attention to those cells in a 2x2 contingency table that correspond to the cause-present cases here than they do to the cells corresponding to the cause-absent cases here (see also Kao & Wasserman, 1993). It is also consistent with other work surveyed by Mandel (this volume) showing that the majority of participants interpret a causal claim such as 'X causes Y' as meaning that when X happens, Y happens. In Mandel's (this volume) terms, such participants have a sufficiency interpretation of causality. It is possible that for the majority of participants in our experiments, judging the strength of the causal relation between an antecedent and a consequent is equivalent to answering a question about the sufficiency of the antecedent for the consequent. Certainly the very weak effect of P(C/not-A) for judgements about causal conditionals suggests that is what most participants are doing when judging the likelihood that a causal conditional statement is true.

Although P(C/not-A) was a weaker predictor of judgements about the causal strength in Experiment 1 than P(C/A), nonetheless it was a significant predictor. It is possible that there are individual differences in the tendency to consider the probability of the effect in the absence of the cause when evaluating causal strength. It is already known that when judging causal conditionals, there are marked individual differences in the tendency to consider the probability of the effect in the presence of the cause (Evans et al. 2003) and that cognitive ability is positively associated with that tendency (Evans, Handley, & Neilens, 2007b). Cognitive ability is also associated with the tendency to underweight the C and D cells of the contingency table (see Stanovich, 1999). Mandel (this volume) in his survey of work on interpretations of causality also shows that there are individual differences. So, whereas the majority of participants in relevant studies (Mandel & Lehman, 1998; Goldvarg & Johnson-Laird, 2001; Mandel, 2003) interpret the statement 'X causes Y' as meaning that when X happens, Y happens, a

minority additionally interpret the statement as meaning that when X doesn't happen, Y doesn't happen. An interesting question for future research is whether consideration of false antecedent cases when assessing the strength of a causal relation or even when explaining what causality is, might also be associated with cognitive ability.

Although in the task we have studied, consideration of cause-absent possibilities is not strongly associated with judgements about causal conditionals, it is associated with judgements about concessive conditionals. Beliefs about false antecedent cases were significant positive predictors of judgements about the truth of even-if statements. That is, if we believe the following statement, 'even if car tax is increased, traffic congestion will get worse', then we are also likely to believe that 'traffic congestion will get worse if car tax is not increased'. Thus, although we have found limited evidence of an association between the assertion of causal conditional claims and consideration of cause-absent cases, we have found an association between the denial of causal claims and the consideration of such cases.

We were not thinking of Mandel's judgement dissociation theory when we designed our experiments, but we are struck by the consistency between the predictions of that theory and our results. The theory predicts that in counterfactually undoing outcomes people will focus on antecedents sufficient to prevent the outcome occurring. When evaluating concessive conditionals, people consider cause-absent cases, which, according to judgement dissociation theory (JDT) are the focus of counterfactuals to undo the outcome. That P(C/not-A) is a positive predictor of the probability of concessives suggests that one of the functions of an even-if conditional is to deny the sufficiency of its antecedent in preventing the outcome.

Although our experiments required participants to make judgements about probabilities, it is worth remembering that the psychological account of the suppositional conditional (see Evans et al. 2005) is that people carry out simulations when assessing the probability of conditional claims. Thus, when we write that consideration of cause-present or cause-absent cases is associated with a particular probability judgement, we mean that participants are temporarily supposing the putative cause specified in the antecedent to be present or to be absent and assessing the ease with which they can imagine the effect specified in the consequent. (The reader is directed to recent accounts of reasoning about conditionals by Evans & Over, 2004 and Johnson-Laird, 2006 for an account of other evidence in the reasoning literature in support of the simulation hypothesis). As mentioned in other contributions to this volume (e.g. Mandel, Hitchcock), ideas in the psychological literature about mental simulation originate with Kahneman & Tversky (1982). Whereas other accounts with debts to the simulation heuristic (see Mandel, this volume) have sought to understand causal claims concerning past events, the suppositional theory has, for the most part, been applied to the study of how people understand causal claims about future events. Although there may be disagreements about what to call the mental simulations in each case (counterfactuals, future hypotheticals), psychologically speaking they are likely to involve very similar processes regardless of whether they concern specific events in the past or classes of

events in the future (see Over et al. 2007). The psychological processes underlying mental simulation are, as yet, quite poorly understood. It seems likely that people use some characteristic of their processing as a surrogate for a probability judgement. One possibility is that ease of imagining a particular consequent given a temporarily supposed antecedent might be substituted for a probability judgement (for more on attribute substitution see Kahneman & Fredrick, 2002). However, the mental processes involved in the acts of imagining and supposing remain to be investigated.

6 Conclusions

In the experiments we described here we included only one condition asking for an evaluation of causal strength and we never asked people to evaluate a past tense counterfactual of the kind most often considered in the psychological literature. Nonetheless, we argue that our experiments have quite a bit to say about the relationship between causal and counterfactual thinking. Our results suggest that the mental simulation of hypothetical possibilities underlies people's evaluation of three different types of claim. As had been suggested by earlier findings in the literature, judgements about causal conditional claims are associated only with simulations where the putative cause is temporarily supposed. Judgements about explicitly causal claims are associated with simulations in which the putative cause is temporarily supposed together with simulations where the cause is supposed to be absent. The evaluation of concessive conditionals, which are often understood to deny a causal relationship (see Byrne, 2005), is similarly associated with judgements based upon both types of simulation. In contrast to causal judgements, however, the likelihood of the effect in the absence of the putative cause is a positive rather than a negative predictor.

These results suggest that evaluating explicit causal claims, or conditional constructions used to deny such claims, cues a process of simulation that involves imagining that the antecedent event is 'undone'. This is akin to the 'undoing' associated with the generation of counterfactual possibilities and results in a belief that contributes to causal judgements in line with covariation and counterfactual accounts of causality. These findings suggest a mechanism through which people get a sense of the strength of causal relations in the world. The notion of a mental simulation is critical in this account as it is to other accounts described in this volume. However, a fuller understanding of the relationship between counterfactuals and causation will require a better theory of the specific processes involved in running such mental simulations.

References

Anderson, J. R. (1990) *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Bennett, J. (1982) 'Even if', *Linguistics and Philosophy* 5: 403–18.

Byrne, R.M.J. (2005) *The Rational Imagination: How People Create Alternatives To Reality*. Cambridge, MA: MIT Press.

Cheng, P. W. (1997) 'From covariation to causation: A causal power theory', *Psychological Review* 104: 367–405.

Declerck, R. & Reed, S. (2001) 'Some truths and non-truths about even-if', *Linguistics* 39: 203–55.

Evans, J. St. B. T., Handley, S.J., Hadjichristidis, C., Thompson, V., Over, D. E., & Bennett, S. (2007a) 'On the basis of belief in causal and diagnostic conditionals', *Quarterly Journal of Experimental Psychology* 60: 635–43.

—— —— & Neilens, H. (2007b) 'Thinking about conditionals: A study of individual differences', *Memory and Cognition* 35: 1772–84.

Evans, J.St.B.T., Handley, S.J., & Over, D.E. (2003) 'Conditionals and conditional probability', *Journal of Experimental Psychology: Learning, Memory and Cognition* 29: 321–55.

—— & Over, D. E. (2004) *If*. Oxford: Oxford University Press.

—— Over, D.E., & Handley, S.J. (2005) 'Suppositions, extensionality and conditionals: a critique of the model theory of Johnson-Laird & Byrne (2002)', *Psychological Review* 112: 1040–52.

Feeney, A. & Handley, S.J. (2000) 'The suppression of "q" card selections: Evidence for deductive inference in Wason's Selection Task', *Quarterly Journal of Experimental Psychology* 53A: 1224–43.

Goldvarg, Y. & Johnson-Laird, P.N. (2001) 'Naïve causality: a mental model theory of causal meaning and reasoning', *Cognitive Science* 25: 565–610.

Handley, S.J., Evans, J. St. B.T., & Thompson, V.A. (2006) 'The negated conditional', *Journal of Experimental Psychology: Learning, Memory and Cognition* 32: 559–69.

—— & Feeney, A. (2004) 'Reasoning and pragmatics: The case of even-if'. In: *Towards an Experimental Pragmatics*, I. Noveck & D. Sperber (eds), pp. 228–53. UK: Palgrave.

—— —— (2007) 'Representation, pragmatics and process in model based reasoning'. In: *The Mental Models Theory of Reasoning: Refinements and Extensions*, W. Schaeken, A. Vandierendonck, W. Schroyens, & G. d'Ydewalle (eds), pp. 25–52. Hove, UK: Lawrence Erlbaum.

Hume, D. (1748) *An Enquiry concerning Human Understanding*.

Jackson, F. (1987) *Conditionals*. Oxford: Blackwell.

Johnson-Laird, P.N. (2006) *How We Reason*. Oxford: Oxford University Press.

Kahneman, D. & Frederick, S. (2002) 'Representativeness revisited: Attribute substitution in intuitive judgment'. In: *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich, D. Griffin, & D. Kahneman (eds.), pp.49–81. New York: Cambridge University Press.

—— & Tversky, A. (1982) 'The simulation heuristic'. In: *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, & A. Tversky (eds), pp. 201–8. New York: Cambridge University Press.

Kao, S.-F. & Wasserman, E.A. (1993) 'Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19: 1363–86.

Konig, E. (1986) 'Conditionals, concessive conditionals and concessives: Areas of contrast, overlap and neutralization'. In: *On Conditionals*, E.C. Traugott, A.T. Meulen, J.S. Reilly, & C.A. Ferguson, pp. 229–46. Cambridge: Cambridge University Press.

Lewis, D. (1973) 'Causation', *Journal of Philosophy* 70: 556–67.

Lycan, W.G. (1991) 'Even and Even If', *Linguistics and Philosophy* 14: 115–50.

Mandel, D.R. (2003) 'Judgment dissociation theory: An analysis of differences in causal, counterfactual, and covariational reasoning', *Journal of Experimental Psychology: General* 137: 419–34.

—— & Lehman, D.R. (1998) 'Integration of contingency information in judgments of cause, covariation, and probability', *Journal of Experimental Psychology: General* 127: 269–85.

Moreno-Rios, S., Garcia-Madruga, J., & Byrne, R.M.J. (2008) 'Semifactual "even if" reasoning', *Acta Psychologica* 128: 197–209.

Over, D.E., Hadjichristidis, C., Evans, J. St. B.T., Handley, S.J., & Sloman, S.A. (2007) 'The probability of causal conditionals', *Cognitive Psychology* 54: 62–97.

Ramsey, F.P. (1931) *The Foundations of Mathematics and Other Logical Essays*. London: Routledge and Kegan Paul.

Sanford, D.H. (1989) *If p then q: Conditionals and the Foundations of Reasoning*. London: Routledge.

Shanks, D.R. (1995) *The Psychology of Associative Learning*. Cambridge: Cambridge University Press.

Stanovich, K.E. (1999) *Who Is Rational?* Mahwah, NJ: Erlabaum.

—— & West, R.F. (1998) 'Individual differences in rational thought', *Journal of Experimental Psychology: General* 127: 161–88.

Wason, P.C. (1965) 'The contexts of plausible denial', *Journal of Verbal Learning and Verbal Behaviour* 4: 7–11.

Author Index

Subject Index