A Reverse Counterfactual Analysis of Causation

DISSERTATION SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE

UNIVERSITY OF CAMBRIDGE

Alex Broadbent

King's College

October 2007

A Reverse Counterfactual Analysis of Causation

Alex Broadbent

Lewis's counterfactual analysis of causation starts with the claim that c causes e if $\sim C > \sim E$, where c and e are events, C and E are the propositions that c and e respectively occur, $\sim$ is negation and $>$ is the counterfactual conditional. The purpose of my project is to provide a counterfactual analysis of causation which departs signigicantly from Lewis's starting point, and thus can hope to solve several stubborn problems for that approach. Whereas Lewis starts with a sufficiency claim, my analysis claims that a certain counterfactual is necessary for causation. I say that, if c causes e, then $\sim E > \sim C$ — I call the latter the *Reverse Counterfactual*. This will often, perhaps always, be a backtracking counterfactual, so two chapters are devoted to defending a conception of counterfactuals which allows backtracking. Thus prepared, I argue that the Reverse Counterfactual is true of causes, but not of mere conditions for an effect. This provides a neat analysis of the principles governing causal selection, which is extended in a discussion of causal transitivity. Standard counterfactual accounts suffer counterexamples from preemption, but I argue that the Reverse Counterfactual has resources to deal neatly with those too. Finally I argue that the Reverse Counterfactual, as a necessary condition on causation, is the most we can hope for: in principle, there can be no counterfactual sufficient condition for causation.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Previously Published Work

Chapter 6 is based on 'Reversing the Counterfactual Analysis of Causation' in *International Journal of Philosophical Studies* (Broadbent, 2007) and appears under the terms of the copyright agreement.

Statement of Length

This thesis contains 78,833 words including footnotes and appendices but excluding Bibliography, and is therefore within the limit of 80,000 set by the Degree Committee of the Department of History and Philosophy of Science.

Acknowledgements

This thesis is the product of guidance and kindness which nobody could claim to have deserved. I count myself extremely lucky to have been supervised by the insightful yet patient Peter Lipton, an extraordinary thinker and teacher. I am also very grateful for kind yet unflinching discussions with my advisor, Martin Kusch. I am grateful to Dan Heard and Torben Rees for putting the proof back into proof-reading — their comments and advice have greatly improved the thesis. Others who have been generous with their time and brains include Arif Ahmed, Alexander Bird, Gunnar Bjornsson, Darren Edge, Ned Hall, Tim Lewens, John McFarlane, Kit Patrick, Arash Pessian, Mark Sprevak, Richard Tuck, Sridhar Venkatapuram, the audience at the Graduate Conference at Bristol in 2005, anonymous referees at the *International Journal of Philosophical Studies* and *British Journal for the Philosophy of Science*, and no doubt other kind people whose omission is my own fault and no reflection of the true extent of my debt. I also owe a lot to some people I have never met, notably David Lewis, to whose work this thesis is in large part an appreciative reaction. The collection *Causation and Counterfactuals* (Collins, Hall and Paul 2004) deserves special mention as a comprehensive and reliable guide to this lumpy terrain. Institutional support has been plentiful. At the University of Cambridge, I am grateful for my Domestic Research Studentship, and to the Board of Graduate Studies (especially Laurie Friday, Kate Maxwell and Kathy White), the Isaac Newton Trust, the Department of History and Philosophy of Science (especially Tamara Hug), the Ferris Travel Fund and King's College, where Maria de la Riva's kindness and competence have been life-savers on more than one occasion. I am grateful to Harvard University for a Visiting Fellowship in autumn 2006, and to the Philosophy Department there. I am indebted in more ways to friends and family than I can hope to justly acknowledge. Particular thanks go to: Knut Nygaard, for pizza and weights in Boston; Julian Hendrix for cat-sitting, house-sitting and wife-sitting while I was at Harvard; Richard Lloyd Morgan, for listening; Jackie Solomon and Maria Whelan, for making my year at the Graduate Union so fun; and my excellent friends at the Cambridge University Powerlifting Club. I thank my family for years of love and support. Finally, I thank my wife, Zeynep: for living so gracefully in recent months with a thesis, as well as a husband; for contributing more to my happiness and endeavours, including this thesis, than anything else; and for her sweetness, fire and love.

Contents

Chapter 1

Introduction

1.1 Two Problems

Suppose I am playing pool, and I miss a shot. My opponent accuses me of being a little tipsy. I deny it vehemently: there was a definite slope on the table. My accuser might say, "Well, slope or not, you must be drunk: otherwise you would never have missed a shot like that."

"Codswallop!" I reply. "It's the slope. If there hadn't been a slope, I would have made the shot, no problem."

"Maybe," replies the silver-tongued scoundrel, "But even if the table does, as you claim, possess some mild tilt, that doesn't change the fact that if you were sober you would have made the shot."

The argument is about causation, even though the word "cause" does not feature. My slanderous companion is applying a common test for causal relevance, formalised in legal contexts as the *sine qua non* or *but for* test. It has the form of a counterfactual: c causes e if, if c hadn't happened, then e wouldn't have happened. My companion claims that if I hadn't been drunk then I wouldn't have missed, which — we both take it — is tantamount to claiming that my drunkenness is the reason I missed. In response, I also apply the test, claiming that if there had been no slope, then I wouldn't have missed. I am thus blaming the slope. But my companion points out that my counter-claim is compatible with his accusation.

This is the first of the problems which I want to address. It is called the *problem of causal selection*. When anything happens, we can often identify a very large number of conditions upon which the happening of that event depends. Many of our ordinary activities demand the presence of oxygen in the atmosphere, for example. However, we very rarely mention oxygen when we are

thinking about the causes of particular events. I know that oxygen is required for me to cook dinner, but I do not say that the oxygen caused the dinner to be cooked. The problem of causal selection is the problem of explicating the difference between the cause (or causes) and the mere conditions for an effect. It is a particularly pressing problem because, although we often rely heavily on the distinction between cause and condition, the counterfactual just mentioned does not respect it. And that counterfactual features in far more serious contexts than the trivial conversation I have just recounted, notably legal contexts.

Now suppose that, after a few more shots from the same locker, I lose the game of pool. "Perhaps you are right," I moan. "If only I hadn't drunk all that beer: then I wouldn't be drunk, and I would have beaten you easily."

"I'm not so sure," says my friend.

"Well, you're a terrible player," I say, with wholly uncharacteristic rudeness. "I would have won if I were sober."

"But that's not what I was questioning," says my friend slyly.

"I don't understand," I admit.

"Well, you said that if you hadn't drunk all that beer, you wouldn't have been drunk. But that's not true. If you hadn't drunk beer, I strongly suspect you would have drunk something else — wine, perhaps. So you would still be a little tipsy, as you usually are by this stage in the evening."

"Maybe, but it was the beer that made me drunk this time," I grumble.

We are exploring a second problem: the *problem of causal redundancy*. Sometimes, causes are redundant: they have back-ups. There are various cases of this sort: in this case, the back-up — my drinking something else — remains potential, and does not actually occur. In other cases, the back-up does occur: for example, an assassin shoots at the President, killing her; at almost the same time, another assassin shoots, and her bullet hits the President just after the first assassin's bullet has killed her. The second assassin's shot does not kill the President, but would have done, had the first assassin not fired.

The problem of redundancy, like the problem of selection, presents difficulties for the *sine qua non* test for causation. However, it presents the opposite kind of difficulty. We saw that the problem of selection threatens the *sufficiency* of the *sine qua non* test: it shows that there are many conditions *sine qua non* which fail to be causes; hence the counterfactual does not suffice for causation. The problem of redundancy, on the other hand, threatens the *necessity* of the *sine qua non* test. It shows that there are cases where, if the

cause had not happened, the effect might nevertheless have happened anyway.

These two problems will be the focus of this project. To give them sharper form, and to understand how they will be approached, it will be useful to agree some background.

1.2 Background

The modern philosophical approach to analysis of causation begins with David Hume [1748]. He brought into focus the way we routinely go beyond what we observe in ascribing causation when all we ever see is coincidence. I strike a match, and it lights; this is all I see, but I further believe that striking the match caused it to light. This prompts two obvious questions. The first is epistemological: How do I know that striking the match caused it to light? The second is conceptual: What do I mean when I say that striking the match caused it to light? When I believe this, what do I believe? These two questions are closely connected: to justify my claim that striking the match caused it to light, I must have some idea what further claims would bear out its truth or show its falsity; and for that, I must have some idea what causation is, or at least what it entails.

Philosophers approach the task of answering these two questions in rather different ways. The approach which is most strongly suggested by my formulation of these Humean questions, and the one I shall take, focuses on the twin tasks of understanding and justifying singular causal claims. Thus I shall not be discussing the meaning of and further justification required for general causal claims, such as "Watching television rots the brain". Nor shall I be concerned with the kind of justification which might refute someone who has serious sceptical doubts about the existence of causation.[1] The epistemology of causation *is* part of the analysis of causation, but only in a rather quiet way: agreeing that causation has certain features might help us to justify, and hence prove the truth of, particular causal claims, if those features can be shown to be present in the particular cases in question. But this sort of justification, although it can lead to increase in knowledge, is not designed to allay sceptical doubts; nor is it suitable as a general description of the way we acquire causal knowledge. Its main purpose is argument: the goal of participation in which, is to link the matter in dispute to other facts which are agreed upon, in such

[1] An analysis in terms whose epistemology is less problematic would obviously help answer a sceptic, but mine shall be a counterfactual analysis, and the epistemology of counterfactuals is no less problematic than the epistemology of causation itself.

a way as to persuade a rational disputant to agree, to jettison agreed facts or to appeal to other facts which are not agreed.

Argument is also the goal of the other twin task, understanding causal claims. Agreeing on the nature of causation allows us to argue about it, in the manner just described. Once an understanding of a causal claim is agreed, it becomes possible to justify it, by arguing that the case in question possesses whatever features our agreed understanding says that cases of causation possess.

Seeking to provide material for disputes, albeit hypothetical ones, is central to conceptual analysis, as I understand it. Sometimes, though, conceptual analysis is considered to be a rather fruitless activity. Thomas Bontly cites a range of reasons found in the literature for losing faith in conceptual analysis: that psychological evidence tells against the notion that concepts have necessary and sufficient conditions, exemplars providing a more accurate model; that Quinean attacks on the analytic/synthetic distinction show there are no purely conceptual truths; and that content externalism in the philosophy of mind raises the prospect that the necessary and sufficient conditions for instantiating a philosophically interesting concept like "causation" might be as much outside the head as the necessary and sufficient conditions for instantiating the concept "water" (Bontly 2006, 178). And, of course, there is no widely accepted conceptual analysis of causation.

This has led some philosophers to pursue other kinds of analysis, notably what Bontly calls *empirical analysis.* Bontly summarises thus:

> ...good old conceptual analysis may simply be the wrong way for reductionism to proceed. Instead, many philosophers today are prone to think of the analysis of causation as an empirical matter, the idea being to reduce causation to some relation uncovered by natural science. Such a reduction would be synthetic, not analytic; a posteriori rather than a priori. And there are several intriguing proposals as to what, on such an account, the causal relation might turn out to be.
>
> (Bontly 2006, 178)

This is not the sort of analysis I shall be pursuing. The reason is that it would not seem to answer the need which I see conceptual analysis as addressing, that is, the argumentative need. Even if it could be established that some physical property was present in all and only cases of causation, this would

not speak in an obvious and direct way to our ordinary disputes. For those disputes concern our ordinary concept of causation. Moreover, it is hard to see how likely candidates for causation could operate both between micro-particles and between the kinds of ordinary things we handle with our ordinary concept: chairs, tables, cars, and so on. This is not to detract from the interest of the project of empirical analysis; but it is a different project, answering different needs, from the one I shall be embarking on.

Nor do I feel forced to abandon conceptual analysis by the kinds of criticism recently mentioned. Maybe our concepts, as psychological entities, are not well modeled as having necessary and sufficient conditions. That does not mean there is no need for conceptual analysis; it might change the shape and tools appropriate to that analysis somewhat, with a greater focus on examples — but modern thinking about causation is already driven by examples. The fact, if it is one, that there are no conceptual truths does not immediately entail that there is no answer about whether a concept is correctly applied in particular cases; and even if it did, the practical pressure to handle those cases (eg. in the courts) forces us to say something about them. Philosophers ought to be well-placed to contribute. And the fact that the meaning of causation might not be "in the head" clearly does not mean that we can't argue about whether causation is present in particular cases, especially when we have not identified a plausible physical candidate for the essence of what we ordinarily call causation, as we have in the case of what we ordinarily call water.

David Lewis builds a conceptual analysis of causation (cf. Collins et al. 2004, 30-39) from the intuitive counterfactual previously described: if c hadn't happened, e wouldn't have happened. He proposes that counterfactual as a sufficient condition for causation. That is, he claims:

> If $\sim C > \sim E$ then c causes e.[2]

Lewis takes the relata of causation to be distinct and actually-occurring events.

Since Lewis makes this counterfactual sufficient for causation (Lewis 2004a, 78), he effectively ignores the first of our problems, the problem of causal selection. For there are many events which meet his condition, and thus suffice for causation according to Lewis, but which we do not ordinarily call causes.

> We have the icy road, the bald tire, the drunk driver, the blind corner, the approaching car, and more. Together, these cause the

[2] c and e are events, C and E stand for the propositions that c and e respectively occur, $\sim$ is negation and $>$ indicates the counterfactual conditional.

> crash... And the crash depends on each. Without any one it would not have happened... But these are by no means all the causes of the crash. ...each of these causes in turn has its causes; and those too are causes of the crash.
>
> (Lewis 1986a, 214)

In my view, this attitude is a departure from the project of conceptual analysis as I understand it to be motivated. It is a departure from the chamber of ordinary debate about causation. For it is a departure from our ordinary concept of causation. Clearly, when we put the claim "c caused e" to ordinary use, we rely on a lot more than the claim that if c hadn't happened, e wouldn't have happened. Otherwise, the pedestrian who is hit by the car that mounts the pavement is as much to blame for her injury as the drunken and reckless driver. The oxygen in the room is as good a predictor of a flame as the striking of the match. My birth is as good an explanation of my late arrrival as the delayed train. And so on.

Lewis accepts that there is more to many of our causal claims than the counterfactual he singles out, but he thinks that this extra can be stripped off, and that underneath is a core notion, free of "principles of invidious discrimination" (Lewis 1973a, 162), which all our causal claims have in common. To make good on this claim, it is necessary not only to argue that there is such a notion, but that the parts which have been stripped off can also be explained. Further, it must be explained why these parts are commonly associated with causation, and why our ordinary application of causal concepts is often *precisely* to make "invidious" discriminations between causes and mere conditions for an effect: as, for instance, when we distinguish the innocence of the pedestrian from the culpability of the driver. — At any rate, so I shall argue.

Much more commonly, it is argued that Lewis's theory does not identify any such core of our causal concept. This argument focuses on the second of our problems, causal redundancy. The problem presents a challenge to Lewis to come up with a necessary condition for causation. For in cases of causal redundancy, it is false that, if the cause had not occurred, the effect would not have occurred. Various efforts have been made to produce more complicated analyses which nevertheless centrally employ counterfactuals. Here, the argumentative burden on me is rather different. I shall not need to argue for the importance of accounting for causal redundancy, as it is widely accepted (unlike the importance I allege for the problem of causal selection). I shall,

therefore, devote my discussion of causal redundancy to arguing for my own counterfactual necessary condition for causation, which remains true in cases of redundancy.

That is the general shape of the project — the backdrop against which it is pursued, and the general aims it has. Let me now give a more detailed outline.

1.3 Outline

Central Claim

My analysis will start from this claim:

> If c causes e then $\sim E > \sim C$.

The counterfactual in this claim is often just Lewis's counterfactual, reversed: so I call it the Reverse Counterfactual. It is offered as a *necessary* condition for causation, whereas Lewis's counterfactual is proposed as a sufficient condition. I shall be defending the Reverse Counterfactual as a necessary condition, only discussing sufficient conditions for causation in Chapter 7, where I shall argue that there is no counterfactual sufficient condition for causation. Thus my analysis is incomplete. However, that does not mean it fails to serve the purposes of conceptual analysis described previously; moreover, since there is a principled reason why no sufficient condition for causation is available, it amounts to the most we can expect of a counterfactual analysis of causation.

Following is an abstract of each remaining chapter.

Chapter 2: Lewis's Asymmetry Thesis

This chapter lays out the elements of Lewis's semantics for counterfactuals which will be relevant for us, then discusses criticisms of one aspect of the theory: the thesis that counterfactual dependence is temporally asymmetric. Lewis's semantic theory is sketched, followed by Fine's objection to the role in that theory of similarity between possible worlds. Lewis presents his fully-fledged position on the asymmetry of counterfactual dependence as a response to Fine's objection. I make an effort to get clear on the proper structure of Lewis's position, particularly on the relation between three alleged asymmetries: the asymmetry of *overdetermination*, the asymmetry of *miracles*, and the headliner asymmetry of *counterfactual dependence.* I endorse three criticisms of Lewis's position. First, I draw attention to the fact that Lewis's

own semantics requires the truth of at least some backtracking counterfactuals, or backtrackers. Second, I introduce *Bennett-worlds*. Third, I evaluate Elga's argument from certain statistical mechanical considerations, and compare *Elga-worlds* to Bennett-worlds.

Chapter 3: Backtrackers Can Be True

The purpose of this chapter is to argue that backtrackers can be true, and to provide a method for working out whether they are true. I argue that backtrackers and foretrackers are not confined to distinct contexts, as Lewis suggests: they are compatible, and can be true in the same context. I argue further that the grammatical complexity with which we prefer to express backtracking reasoning need not indicate a relevant shift of context. Whether an asymmetry of counterfactual dependence exists and what its nature might be is, therefore, an open question, not to be settled by appeal to grammar. Against that background, I suggest that we need a method to help us tell when backtrackers are true, irrespective of distracting linguistic features. This amounts to a further criticism of Lewis's asymmetry of counterfactual dependence, because it deprives that thesis of its explanandum — an ordinary context in which the resolution of counterfactual vagueness makes backtrackers mostly false. Lewis's "visualisation" method is criticised, and the Ramsey Test is found unsuitable as it stands. Drawing on hints from Frank Ramsey, Nelson Goodman and Dorothy Edgington, I propose a simple test which seeks to harness our robust and temporally neutral abilities to make and assess inductive inferences for the task of assessing counterfactuals. I dub it the *Inference Test*.

Chapter 4: Selection

I strike a match, and it lights. It is unusual to say that the presence of oxygen caused the flame, even though we might be fully aware that oxygen is needed for the flame. This chapter presses the problem of selecting the cause from among mere conditions, which has often been dismissed by theorists of causation as a sort of whim. I start by arguing that selection needs to be accounted for, and sketching some of the approaches which have been tried. I focus on the strategy of assimilating causal selection to the contrastive mechanism of causal explanation. Although this approach enjoys some success, I argue that it fails either properly to explain or to enable us to justify our selective

practices. I introduce the *Reverse Counterfactual* which, I suggest, captures the intuitive notion that causes *make the difference* to their effects. I argue that the Reverse Counterfactual is true of causes but not of mere conditions. This yields an account of selection which overcomes the objections raised against the contrastive strategy, and which links the context-sensitivity of causal selection to the context-sensitivity of counterfactuals. The relation between contrastive explanation and the Reverse Counterfactual is discussed, along with various objections.

Chapter 5: Transitivity

This chapter argues that the causes of the cause of an effect are not always causes of that effect: that is, causation is not transitive. I begin by distinguishing various kinds of nontransitivity. I consider four motivations for the common view that causation is transitive, and argue that none is compelling. Two sorts of counterexamples to transitivity are then distinguished: those due to *distance* (or failure of *proximity*) and those due to a special causal structure, *double prevention*. I argue that proximity failure is closely linked to causal selection. The Reverse Counterfactual tends to support the intuitive view that very distant events are not causes, even if they may be mere conditions. Then I argue, contrary to some recent literature, that cases of double-prevention fail to be cases of causation. The Reverse Counterfactual agrees. Efforts by McDermott, Paul and Hall to diagnose the difficulty with double prevention are considered and found wanting. However, I agree with these authors that double prevention does not constitute a real counterexample to causal transitivity; I suggest that only an unselective notion of causation could persuade us to think otherwise. Finally I consider how a valid substitute for counterfactual transitivity might help explain why causation sometimes appears to be transitive (unlike other nontransitive relations such as touching), and also why distance should be relevant to causal transitivity.

Chapter 6: Redundancy

The term "redundant causation" is an umbrella. Overdetermination among existing events is distinguished from redundancy due to non-occurring backups. Symmetric and asymmetric overdetermination are further distinguished, and within the latter, preemption and trumping are distinguished. It is proposed that the Reverse Counterfactual holds true of cause-effect pairs even in

cases of preemption. This gives us the basis for distinguishing causes from preempted events. I argue that the Reverse Counterfactual is true only of causes in some easy cases, allowing us to distinguish causes from preempted events. Then I argue that we can make the same distinction in harder cases, if we employ an intuitive notion of a causal chain. Next we discuss trumping, where appeal to causal chains appears not to help. I argue that the Reverse Counterfactual distinguishes trumping events from trumped events with whatever intuitive resources are provided to motivate the intuitive distinction; if no such resources are provided, trumping collapses into symmetric overdetermination. Finally we discuss symmetric overdetermination, which does not yield counterexamples in the way that preemption does. Nevertheless I argue that the Reverse Counterfactual account avoids the difficulties which beset Lewis's account associated with mereological summing.

Chapter 7: Conclusion

A synopsis of the defence of the Reverse Counterfactual as a necessary condition for causation is presented. I then argue that there can be no counterfactual sufficient condition for causation. I argue that simultaneous causation occurs. Since the asymmetry of counterfactual dependence is temporal, counterfactual dependence can never offer a full characterisation of causation, no matter how sophisticated a thesis might be proposed of the asymmetry of counterfactual dependence. Thus there is more to causation than counterfactual dependence among particular events. I finish by compiling the more complicated versions of the Reverse Counterfactual which are developed in earlier chapters, along with other important claims which have been defended, in a concise summary.

1.4 Other Problems

There are various problems which I do not intend to tackle. Three stand out: the nature of causal relata, the question whether causation is a relation, and issues concerning probabilistic causation.

Lewis takes the causal relata to be events, which he further takes to be properties of spatiotemporal regions (Lewis 1986c). I do not think that we need total clarity on the nature of the causal relata, before we can make progress with a conceptual analysis of causation. Even if causation is not between events, or if events are not what Lewis says they are, it is to be hoped that

a substantive analysis of our ordinary concept, such as we might appeal to in settling a dispute, will depend only lightly if at all on the details.

I also try to avoid committing myself on whether causation is a relation. When discussing causal transitivity and asymmetry, I take these properties of causation to obey the logic of relations. Since causation must be relevantly like a relation to support these properties, I treat causation as a relation for the purposes of those discussions. However my analysis supports causation among absences, which is potentially problematic for the view that causation is a relation, since relations need relata. I leave this tension unresolved. Whether or not causation is a relation, it seems that it should be able to support transitivity and asymmetry. On the other hand, absences feature as causal "relata" in ordinary talk, and thus it seems to me that a theory of causation must also feature them if it is correctly to capture the nature and extension of the concept underlying our ordinary talk.

If this tension is resolved with some clever way in which causation can have the relational properties of asymmetry and transitivity, yet still meaningfully admit causal absences, then my theory will meet with no problem. If, however, absences are denied as causal relata, then not only will my theory be in error, but so will our ordinary causal talk. In that case, the theory is none the worse as a conceptual analysis. Whatever directive is given for excluding absences and replacing them with presences will apply equally to the ordinary talk and to the theory: it may simply be appended to the theory. For these reasons, although it is an interesting question and important in other ways, I do not think it matters to be precise about whether causation is a relation when we are trying to provide a substantive analysis.

Perhaps most controversially, I eschew discussion of probabilistic causation and of causation under indeterminism. There is little doubt that our ordinary talk about causation often overlaps with talk about chance and probability. We might say that smoking causes lung cancer, although it does not render lung cancer inevitable, merely more probable. An analysis of the ordinary concept of causation must therefore say something about how causation relates to these concepts, even if it completely distinguishes them. I agree with Salmon that simply denying any relation between causation and probabilistic concepts is unsatisfactory.

> It may be maintained, of course, that in all such [probabilistic] cases a fully detailed account would furnish invariable cause-effect relations, but this claim would amount to no more than a declaration

> of faith. ...it is as pointless as it is unjustified.
>
> (Salmon 1993, 137)

So I do not want to dismiss discussion of probability and chance out of hand.[3]

On the other hand, I do not want to discuss them, either. I might slightly justify setting aside probabilistic causation in order to divide and conquer. While exploring the relation between counterfactuals and causation, it might make sense to ignore other difficult issues, even if they bear directly on the subject of analysis — just as a physicist might ignore friction when studying the laws of motion. Such an attitude can never be permanent, because the phenomena which are idealised away must be included in any final analysis. But the attitude may be acceptable in the interim, when no final analysis exists; and that, surely, is the situation with the conceptual analysis of causation.[4]

1.5 Thinking About The Reverse Counterfactual

Finally before we begin, it must regretfully be admitted that the central idea is difficult. It is not complex: it is quite simple to say that, in general, if the effect hadn't happened then the cause wouldn't have happened. But in some ways simplicity makes the claim harder to explicate. Even though I am convinced this claim is true, I fully admit that we do not normally say so, or think so. One way to ease the headache, which has already been mentioned, is to think of it as a *counterfactual sufficiency* claim. On Lewis's account, causes are *counterfactually necessary* for their effects; but I am proposing that they are *counterfactually sufficient* (albeit in a slightly subtle sense, elaborated in 6.1). Another, perhaps more intuitive, pill to ease the headache is the idea of *difference-making*, discussed in Chapter 4. To say that c causes e is, plausibly (and in the absence of overdetermination), to say that c makes the difference between e occurring and e not occurring. One way to understand the claim that c "makes the difference" is that c just *is* the difference between the case where e occurs and the relevant counterfactual scenario where e does not. But if c is

[3] Contributors to the compatibilism debates concerning, on the one hand causation and indeterminism, and on the other hand and determinism and chance, include: Anscombe 1971, von Wright 1974, Lewis 1986a, 1994, Loewer 2001, Dowe and Noordhof 2004, Schaffer 2007.

[4] Collins, Hall and Paul say that a divide-and-conquer methodology is acceptable, given how tough the analysis of causation is (Collins et al. 2004, 38–39).

the difference between e occurring and not, then if e had not occurred, c would not have occurred. (Otherwise c would not after all be any difference between e occurring and not.) It therefore seems that the counterfactual, which is difficult to get one's head around, follows by an intuitive (if not watertight) line of reasoning from a plausible understanding of at least some causal situations: that causes "make the difference" to their effects. This conception of difference-making may therefore be brought to mind in many cases where the Reverse Counterfactual is difficult to comprehend or assess.

Chapter 2

Lewis's Asymmetry Thesis

2.0 Abstract

This chapter lays out the elements of Lewis's semantics for counterfactuals which will be relevant for us, then discusses criticisms of one aspect of the theory: the thesis that counterfactual dependence is temporally asymmetric. Lewis's semantic theory is sketched, followed by Fine's objection to the role in that theory of similarity between possible worlds. Lewis presents his fully-fledged position on the asymmetry of counterfactual dependence as a response to Fine's objection. I make an effort to get clear on the proper structure of Lewis's position, particularly on the relation between three alleged asymmetries: the asymmetry of *overdetermination*, the asymmetry of *miracles*, and the headliner asymmetry of *counterfactual dependence.* I endorse three criticisms of Lewis's position. First, I draw attention to the fact that Lewis's own semantics requires the truth of at least some backtracking counterfactuals, or backtrackers. Second, I introduce *Bennett-worlds.* Third, I evaluate Elga's argument from certain statistical mechanical considerations, and compare *Elga-worlds* to Bennett-worlds.

2.1 Introducing Counterfactuals

What I call counterfactuals are also known as counterfactual conditionals or subjunctive conditionals, and are usually considered to be typified by sentences of the form,

> If it were the case that A then it would be the case that C.

We can also write the above as $A > C$.

We will discuss Lewis's conception of counterfactuals shortly, but at least some properties of counterfactuals are widely agreed upon. Specifically:

- counterfactuals do not imply the truth of their antecedent, either logically or by convention in ordinary use (a property they share with other conditionals);
- counterfactuals are not true in virtue of the falsity of their antecedent (making them distinct from material conditionals);
- counterfactuals are not true in virtue of the truth of their consequent (again, distinguishing them from material conditionals).

In other words, counterfactuals concern not just what *is* the case but what *would be* the case. Analysing counterfactuals comes down largely to explaining what that means.

The word "counterfactual" is used in two other ways, apart from this use as a noun. It is used as an adjective to mean simply contrary to the facts. It is also used in the phrase "counterfactual dependence". I shall take Lewis's definition of counterfactual dependence. So it is something which holds between events, and c counterfactually depends on a iff it is the case both that $A > C$ and that $\sim A > \sim C$ (Lewis 1973a, 166–7).[1]

The temporal direction of counterfactuals will be central here. What I shall call a foretracking counterfactual, or foretracker, is one which goes forwards in time: the antecedent denotes matters at a time earlier than the time of what the consequent denotes. If the example above is a foretracker, a is before c. A backtracking counterfactual, or backtracker, is one which goes backwards in time: the antecedent denotes matters at a time later than the time of what the consequent denotes. If the example above is a backtracker, c is before a. We might try to be precise about what it is for a to be before c: whether, in particular, a must be wholly before c, or must start before c starts, or must end before c ends. For present purposes, however, I see no advantage in precision — we only risk ruling out intuitive cases for no benefit, since my theory will not rely on the notion of beforeness (even if Lewis's does). Moreover, in the

[1] At the point to which the reference refers, Lewis appears to define counterfactual dependence among propositions rather than events. But it is more natural to see the dependence as being between events: after all, it is surely something about the events a and c, not the propositions that they occur, which gives rise to the facts that $A > C$ and $\sim A > \sim C$. Elsewhere Lewis himself speaks of counterfactual dependence between events, for instance in the introduction to his *Philosophical Papers — Volume II* (Lewis 1986a, xii).

cases we shall consider, it will be obvious what temporal direction, if any, the counterfactual in question has.

One reason that Lewis's counterfactual account received so much attention is that it was based firmly in a comprehensive semantic theory of counterfactuals. The theory of causation benefited in a general way from the background theory of counterfactuals, because the background theory eased the worry that to analyse causation in terms of counterfactuals is merely to swap one mystery for another. The theory of causation also benefited more directly, since — as we shall see — it makes direct use of some of the properties which Lewis's semantics ascribes to counterfactuals. In particular, Lewis's theory of causation relies upon his thesis of the asymmetry of counterfactual dependence, which can be roughly characterised as a ban on backtrackers. The analysis of causation which I will propose employs backtrackers, so I must argue not only against Lewis's theory of causation, but also against his semantics for counterfactuals, at least insofar as it implies the falsity of the counterfactuals I wish to use to analyse causation.

Before we turn to that, an issue should be set aside. It may be doubted whether counterfactuals — and indeed conditionals more generally — should be considered to bear truth-values. A non-factualist such as Richard Bradley [2007] would deny it, and would thus take believing a counterfactual to be something other than believing it to be true. I will assume that counterfactuals can and do bear one of two truth-values, true or false, and further that when we believe them, we believe them to be true or false. This is not because I am hostile to a non-factualist view. It is rather that I do not think the sorts of changes which a non-factualist view would require will ultimately affect the prospects for a counterfactual analysis of causation. The non-factualist does not assert that there is no difference between those counterfactuals I call true and those I call false, only that the difference is not a difference of truth-value. Whatever the difference is, I hope it could be plugged into a counterfactual analysis of causation, and preserve the substance of that analysis.

2.2 Lewis's Semantics for Counterfactuals

2.2.1 Possible World Semantics

We have given a working definition of a counterfactual as something typified by this sentence-type:

> If it were the case that A, then it would be the case that C.

Over-attention to surface grammar can be unproductive, however. Sometimes we might express ourselves differently, but mean much the same thing. At other times we might express ourselves with this sort of construction, but use it in an unusual way. Both difficulties may be compounded if there are languages into which the English subjunctive construction does not straightforwardly translate. We can avoid these difficulties if we identify counterfactuals, not by natural-language expressions of a given form, but by the semantics appropriate for them. It may be that other expressions are best interpreted with this semantics on certain occasions, and that expressions of this form are sometimes better interpreted differently.[2]

Lewis's semantics appeals to possible worlds arranged in a system of spheres of accessibility. The modal logical concepts of spheres and accessibility need not be discussed here. Truth conditions of counterfactuals are given in terms of comparative closeness of possible worlds, where closeness is a weak ordering (cf. Lewis 1973b, 13–19). Lewis proposes that closeness is a relation of comparative similarity (Lewis 1973b, 1).

The proposed truth conditions for counterfactuals are summarized as follows.

> In brief: a counterfactual is vacuously true if there is no antecedent-permitting sphere, non-vacuously true if there is some antecedent-permitting sphere in which the consequent holds at every antecedent world, and false otherwise.
>
> (Lewis 1973b, 16)

Roughly speaking, $A > C$ is vacuously true if its antecedent is necessarily false, and non-vacuously true if, moving in towards the actual world, we eventually get to a point beyond which all the worlds remaining between us and the actual world where A is true are also worlds where C is true. It is false if we never reach such a point: that is, if, no matter what A-world we take, there is always one (which could be the one we took) as close or closer where $\sim C$.

So far we have only discussed the "would" counterfactual: if it were the case that A, then it *would* be the case that C. Sometimes, however, we also

[2]We can remain neutral, at this point, on what a counterfactual is to be identified *with*. Counterfactuals could still be natural language constructions, but those satisfying certain semantic constraints, rather than having a certain surface form. Or counterfactuals could be identified with the propositions expressed by such statements, or with something else again.

use "might" in counterfactual expressions: if it were the case that A, then it *might* be the case that C. Let us use $\geq$ to stand for the might-counterfactual, in the way we are using $>$ for the would-counterfactual. Lewis interdefines the might-counterfactual and the would-counterfactual, as follows:

$$A \geq C =^{\mathbf{df}} \sim (A >\sim C)$$
$$A > C =^{\mathbf{df}} \sim (A \geq\sim C)$$

(Lewis 1973b, 2 — my symbolism)

The motivation for this definition is as follows. Suppose Jack asserts that if he were to climb the hill, he would fall down and break his crown. Jill denies it, yet she does not assert that if Jack were to climb the hill, he would not fall down. She knows the path is slippery, and that falling down is a real possibility. Rather, when she denies Jack's assertion, she is committing herself to the claim that if he were to climb the hill he *might not* fall down: it is not a foregone conclusion that he would. This is compatible with the claim that he might fall down.[3] To summarise the truth conditions of a might-counterfactual at a world i directly in possible world terms, as opposed to by interdefinition with the would-counterfactual:

> ...the 'might' counterfactual is then true if and only if, as we take smaller and smaller antecedent-permitting spheres around i without end, and thereby confine our attention to antecedent-worlds closer and closer to i, we never leave behind all the antecedent-worlds where the consequent holds.
>
> (Lewis 1973b, 21)

It should also be noted that *would* entails *might*, on Lewis's view: $A > C \vdash A \geq C$. This is roughly because, if some A-world where C is *closer* than any where $\sim C$ (and so $A > C$), it follows that some A-world where C is *at least as close* as any where $\sim C$ (and so, roughly, $A \geq C$).[4]

[3]This definition of the might-counterfactual is thus connected with the denial of excluded middle for would-counterfactuals. Robert Stalnaker disputes both Lewis's denial of excluded middle, and his definition of "might" [Stalnaker, 1981], which he points out is also used in non-conditional contexts. To pursue these niceties would take us too far from the main thread; we will, therefore, set them aside, and accept Lewis's "might", and his rejection of counterfactual excluded middle. (For further discussion, see Bennett 2003.)

[4]In fact the situation is more complicated. Lewis claims that when A is not entertainable, and hence $A > C$ is non-vacuously true, then $A \geq C$ is false (Lewis 1973b, 23–4). I am not sure how this squares with his assertion that $A > C$ and $A \geq C$ are both true when A is impossible (Lewis 1973b, 24–6), assuming that an impossible antecedent fails to be entertainable. It does not matter for our purposes since we shall only be considering non-vacuous counterfactuals of either sort.

In general, when I say "counterfactual" without qualification, I am referring to the would-counterfactual. The might-counterfactual will be explicitly specified when it is meant.

Lewis takes a stand on a couple of points which will be relevant but which are not obvious. First, he holds that counterfactuals whose antecedent and consequent are true at the actual world are true. This is slightly counterintuitive, because normally we would not assert a counterfactual unless we were uncertain of the truth-value of the antecedent, or else thought it false. Serving up a lasagne at my house, I would be unlikely to turn to you and say, "If you were to come round for dinner, I would cook you a lasagne". If I did say it, you would probably take me to be referring to some *other* time, whose occurrence is as yet a mere possibility. But Lewis thinks this is mere conversational, not logical, implication.

This view arises from the *Centering Assumption*, which states that our world is closer to itself than any other world is to ours. Lewis accepts the Centering Assumption (Lewis 1973b, 26–31), for reasons which will become clear in the next section, and this explains why counterfactuals do not imply the falsity of their antecedents.

Another substantive point on which Lewis takes a stand concerns the ordering of worlds. It is sometimes easier to think of $A > C$ as meaning that it is the case that C at the closest world where it is the case that A. But counterfactuals may be true where there is no closest world. There may be either more or less than one closest A-world. If there are several worlds tied for closeness, there is more than one closest A-world. Ties are impossible on the *Uniqueness Assumption*, that there is at most one closest A-world. The Uniqueness Assumption enjoys little support from our ordinary use of counterfactuals, where we tend to specify counterfactual situations rather vaguely, only up to the level of precision we feel to be required. Lewis rejects the Uniqueness Assumption. But there may also be less than one closest world; and the corresponding assumption that there is at least one closest world — the *Limit Assumption* — is intuitively easy to make. The Limit Assumption denies that there may be an infinite series of worlds, each closer to actuality than the last, but none closest (Lewis 1973b, 20). Lewis also rejects the Limit Assumption.[5] His position on both the Limit Assumption and the Uniqueness

[5]Stalnaker endorses both the Uniqueness and Limit Assumptions (Stalnaker 1981, 88-9), arguing that there is exactly one closest A-world. Note that the two assumptions together entail counterfactual excluded middle, if the logical law of excluded middle holds: for there will be exactly one closest A-world, and by logical excluded middle it will be the case that

Assumption are explained by his view of what it is for one world to be closer to actuality than another, to which we now turn.

2.2.2 Closeness As Similarity

Evidently, the sense of Lewis's account of counterfactuals depends on the concept of relative closeness among worlds. Lewis suggests that closeness is a relation of "overall similarity between worlds" (Lewis 1973b, 14). Similarity does an incredible amount of work for Lewis, and certain features are important. "Overall" means that we take the whole possible world into account, not just a part of it. Yet we do not have to take every *respect* of possible difference into account in assessing overall similarity. Some respects of similarity count for very little: for example, the similarities between grue things are of rather little importance (cf. Lewis 1979, 42). Similarity is something which Lewis thinks we have an intuitive grasp on, for instance when we remark on a pair of identical twins or compare two essays for signs of plagiarism.

The more similar a world is, the closer it is. The identification of modal closeness with similarity explains Lewis's rejection of the Limit Assumption. Lewis points out that there is no shortest line longer than an inch, and argues on this basis that there is in principle no closest world in which a given line is longer than an inch, since in principle there is no minimal difference between a line one inch long and a line longer than one inch (Lewis 1973b, 20–21). His adoption of the Centering Assumption can also be explained by his identification of closeness with similarity, since it makes good intuitive sense to deny that any distinct y can be as similar to some x as x itself.[6] The Centering Assumption is sometimes implicitly rejected when philosophers speak of *close* worlds, rather than of some worlds being clos*er* than others. However it will be accepted here, because once again the issues will lead us too far astray, and also in this case because no semantics in terms of non-comparative closeness has gained wide currency.[7]

$C \vee \sim C$ at that world. So $(A > C) \vee (A >\sim C)$.

[6] Stalnaker agrees with Lewis about Centering: "Various constraints are placed on the selection function... For example, it is required that the world selected relative to proposition A be an A-world... And if the actual world meets this condition, it is required that it be selected" (Stalnaker 1981, 88).

[7] In spite of this, it is remarkably common to hear people speak of close possible worlds in a non-comparative sense. A notable example of a denial of Centering doing theoretical work occurs in Robert Nozick's discussion of knowledge. Among his conditions for knowledge, Nozick requires of S to know that p that if p were true, S would believe that p — that is, $p > B_s p$ (Nozick 1981, 178). When an agent holds a true belief, this condition is trivially satisfied on Lewis's semantics. Nozick, however, requires that the belief be true not just at

Kit Fine, among others, puts a simple and famous objection to Lewis's identification of closeness with similarity (Fine 1975). If Nixon had pressed the button, there would have been a nuclear holocaust. But a world with a nuclear holocaust is not very similar to the world we know and love today. There are worlds where the button malfunctions, or the current somehow fails to be transmitted, or something, and by the time Nixon cottons on, he has changed his mind and the cataclysm is averted. These worlds seem, at first sight, to be more similar to ours than any apocalyptic world, overall. So it appears that Lewis ought to deny that if Nixon had pressed the button there would have been a nuclear holocaust. By extension, it seems he must deny any counterfactual where the consequent takes us drastically away from actuality. But ordinarily, we do not assess counterfactuals by how drastic their consequents are: indeed, we sometimes use counterfactuals when reasoning to *avoid* drastic consequences.

This objection prompts Lewis to say more about similarity. When deciding whether one world is more similar to ours than another, we should heed the following four directives. Determinism is assumed in both temporal directions.

> (1) It is of the first importance to avoid big, widespread, diverse violations of law.
>
> (2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
>
> (3) It is of the third importance to avoid even small, localized, simple violations of law.
>
> (4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.
>
> (Lewis 1979, 47-8)

A violation of law is a case where what goes on at the world under consideration contravenes the natural laws of our world:

---

the actual worlds but at relevant close possible worlds. In requiring this, he clearly rejects the Centering Assumption. If Timothy Williamson's safety condition on knowledge is expressed as the counterfactual $B_s p > p$, it also requires the violation of the Centering Assumption, although Williamson denies that he is committed to a counterfactual expression (Williamson 2000, 148–50).

> ...the violated laws are not the laws of the same worlds where they are violated. That is impossible; whatever else a law may be, it is at least an exceptionless regularity.
>
> (Lewis 1979, 44-5)

At our world, w_0, Nixon does not actually push the button. Lewis distinguishes four kinds of worlds where he does. Worlds like w_1 are exactly like ours until shortly before the time of the antecedent, when by a small miracle the antecedent comes true. After that, they follow our laws. w_2 typifies worlds where Nixon presses the button, but no laws are violated: history is suitably different to yield this button push nomically. w_3 is a world where exact match prevails until a small miracle, as in w_1; but unlike w_1, there is a further miracle to suppress the most obvious lawful consequences of the button-push, so that approximate match with w_0 resumes soon after. Finally, worlds like w_4 are like w_3 except that the second miracle secures exact rather than approximate match with w_0.

The closest worlds are typified by w_1, which exhibits an exact match with the actual world in matters of particular fact, up until, or until shortly before, the time of the antecedent. There will be a minor violation of law — a small miracle — shortly before Nixon presses the button. After that, matters unfold lawfully, by the laws of our world. So the missiles are fired, Soviet warning systems detect them and retaliate with an equally devastating salvo, and the surface of Earth is substantially altered and rendered lifeless in a series of huge nuclear explosions. w_1 satisfies (1) totally, (2) partially, and (3) and (4) not at all.

w_1 is closer than worlds like w_2 where Nixon presses the button but no laws are violated. By determinism, w_2 will differ in some degree from the actual world w_0 at every time — future and past — in matters of particular fact. (Lewis argues that the difference will probably be great, because of the way one thing leads to another.) So w_2 violates (2) to a much greater extent than w_1.

Worlds such as w_3 are ones where exact match prevails until just before the miraculous button press, and then, by another tiny miracle, destruction is averted. But w_3 is further from w_0 than w_1. For although w_3 may be similar to w_0 after the button press, it will not be quite the same. Nixon's finger will have left a print on the button. The bottle of gin on the side will be a bit emptier, because he had a swig before he pressed. The man on the moon, watching through his X-ray telescope, will have seen the button press; and

other light waves carrying the same information out of the Pentagon windows will be embarking on their journey to Alpha Centauri. Nixon's memoirs will be different. And so on. So the second small miracle — say, the current dying in the wire — will secure only approximate similarity, (4), which is of less importance than avoiding small miracles, (3). Both w_1 and w_3 violate (3), but w_3 does so twice. So w_1 is more similar to our world, w_0.

Finally, consider worlds like w_4 where by some miracle exact match rather than approximate match resumes after the button press. w_4 is also more remote than w_1. For the miracle needed to suppress all the traces of the button press, and thus secure exact match, will be a huge one. Light waves vanish, heat is sucked back out of the wire, finger prints dissolve, gin vanishes from Nixon's stomach and reappears in the bottle, and Nixon's memories are altered. This is a violation of (1).

There are more arguments against, and counterexamples to, Lewis's four-fold definition of similarity than we will survey here. (For a concise list, see Schaffer 2004b.) In particular, I wish to set aside a whole class of counterexamples seeking to derive advantage from Lewis's assumption of determinism and the difficulty of extending the account to chances. We will be concerned only with direct challenges to the temporal asymmetry which arises from Lewis's response to Fine, and which I shall now lay out.

2.3 Overdetermination and Miracles

2.3.1 The Structure of the Argument

Lewis's asymmetry thesis and the issues surrounding it are complex, and it will help a precise statement of the thesis to work up to it in a logical way. The foundation is another asymmetry, the asymmetry of overdetermination. This gives rise to the asymmetry of miracles, which in turn gives rise to the asymmetry of counterfactual dependence.

This is the exact reverse of the order in which Lewis presents the asymmetries (cf. Lewis 1979, 32-52). Lewis starts by suggesting that counterfactual dependence is asymmetric, for entirely independent reasons which we will consider in the next chapter. He then gives the four-fold definition of similarity which we have just discussed. It takes a much smaller miracle, he argues, to make a world like ours in the past diverge in the future, than it does to make a world unlike ours in the past converge (or reconverge) at a later time. He labels the asymmetry: the asymmetry of miracles. Finally he presents his ex-

planation: the asymmetry of overdetermination. His reason for this order of presentation is my reason for presenting them in the reverse order: that the order of explanation runs from overdetermination, to miracles, to counterfactual dependence. Lewis is proposing his thesis as an explanation and characterisation of various vaguely-felt asymmetries, and consequently he pushes through to deeper levels of explanation as he progresses. In this chapter, however, we are mainly interested in assessing the internal merits of the position, rather than its explanatory virtues.

2.3.2 The Asymmetry of Overdetermination

Throw a stone into a pond. Given the laws of nature, and some background assumptions, you will in theory be able to predict when the waves will start arriving at the shore where you are standing. There is nothing special about where you are standing: you could make this prediction equally for any other point on the pond's circumference. At least, you could if you knew enough maths, physics and facts about the pond and the entrance of the stone, assuming of course that the process is wholly deterministic. Likewise, you could in principle deduce when a projectile entered a pond, by considering the time at which waves start arriving and the curvature of the wave-front, along with various other facts, principles and background assumptions. Again, you need not be standing just where you happen to be standing in order to do this: other points on the shore would do just as well. There are very many wave-segments we could consider, each of which determines the time and location of the point at which the stone struck the water. On the other hand, that event — the stone striking the water — determines very many different wave segments. This process displays an asymmetry: an earlier state of affairs determines very many later ones, and very many later states of affairs determine this one earlier one.[8]

At any rate, this is roughly the picture Lewis wants us to accept. The general thesis of the asymmetry of overdetermination then comes down to the claim that processes like this one are very much more common than processes displaying the opposite characteristics — processes such as stones being propelled from ponds by converging water-waves. Such processes are physically possible, assuming that the fundamental dynamical laws are time-reversible. But they occur very rarely. This brings out the contingency of Lewis's thesis.

[8] According to both Lewis [1979] and Price [1992a], this example originates with Popper [1956].

It is a claim about how our world *happens* to be, not about how it *has* to be.

Let us make the notion of overdetermination precise. *Logical overdetermination* might be thought of as follows. Consider two arguments:

> All men are mortal, and Socrates is a man, hence Socrates is mortal.
>
> Only gods are immortal, and Socrates is not a god, hence Socrates is mortal.

These combinations of premises each entail, and in that sense determine, that Socrates is mortal. But they are not equivalent — neither pair of premises entails either of the other premises.[9] Moreover each premise is ineliminable: remove a premise from one of the arguments, and that argument is no longer valid. As they stand, however, both arguments are sound. The fact that Socrates is mortal might be said to be *logically overdetermined* by the two arguments: each determines it by logical implication; they are not equivalent; and since neither features any redundant premises, neither contains a premise identical with, equivalent to or entailing the premise set of the other argument.[10] (Note that this does not prevent overdetermining premise sets sharing some premises.) More generally we can say that the fact denoted by the proposition C is logically overdetermined by the premise set $\{X_1, ..., X_n\}$ iff:

(i) $\{X_1, ..., X_n\}$ has more than one non-equivalent subset of premises whose conjunctions P, Q, etc, each entail C [an entailment requirement];

(ii) no proper subset of the conjuncts of P, Q, etc, entails the fact C [an ineliminability requirement];

(iii) C does not entail any subset of $\{X_1, ..., X_n\}$ [a non-premise-circularity requirement].[11]

[9] Obviously, the first pair does not entail that only gods are immortal: even though men aren't, other things might be. Nor does it entail that Socrates is not a god: Socrates might be a man and a god. Likewise, the second pair obviously does not entail anything about the mortality of all men, since it does not rule out the possibility that some men are gods. Nor does the second pair entail that Socrates is a man: he could equally be a goat, for example.

[10] To see this, suppose that we have two premise-sets $\{X_1, ..., X_n\}$ and $\{Y_1, ..., Y_m\}$, both of which entail some conclusion C, and that one premise set $\{X_1, ..., X_n\}$ contains a set of propositions $\{P\}$ which entails the other premise set $\{Y_1, ..., Y_m\}$. Then all the other propositions in $\{X_1, ..., X_n\}$ would be redundant for concluding C, since $\{Y_1, ..., Y_m\}$ entails C and $\{P\}$ entails $\{Y_1, ..., Y_m\}$.

[11] If premise circularity is allowed then, since every fact implies itself, a fact would be overdetermined iff there were at least one valid non-circular argument for it.

We can say that C is logically overdetermined in general, without reference to some particular premise set, iff there exists some premise set by which it is logically overdetermined.

Using this notion of logical overdetermination, we can define the perhaps more interesting notion of *nomic overdetermination*: an event e is nomically overdetermined by a set of simultaneous events $\{e_1, ..., e_n\}$ not including e when it is logically overdetermined by the set of propositions $\{E_1, ..., E_n\}$ that those events occur, in conjunction with some law(s) of nature L. (There is no need to require that all the laws of nature are invoked, but to deserve the title "nomic", at least one must be.) That is, e is nomically overdetermined by simultaneous events $\{e_1, ..., e_n\}$ iff it is logically overdetermined by $\{E_1, ..., E_n, L\}$. We can say e is nomically overdetermined without reference to any particular set of events when there exists some set of simultaneous events by which it is nomically overdetermined. We can also say that e is nomically *determined* (rather than overdetermined) by a set of events $\{e_1, ..., e_n\}$ not including e when that set logically determines it in conjunction with the laws.

It appears that Lewis's asymmetry thesis is supposed to apply to nomic overdetermination as we have defined it. In a deterministic world, every fact will have at least one minimal set of conditions which nomically entails it (Lewis 1979, 49). Lewis's claim is that many more such minimal sets lie in the *future*, relative to a given fact, than in the *past.* He gives an example like the one we began with, of a wave emitted from a point source (Lewis 1979, 50), as a special case: many disjoint wave segments determine properties of the source, together overdetermining it. Since the source does not overdetermine the wave segments, there is an asymmetry of overdetermination in such cases, Lewis claims. And there are many other cases like that, not confined to wave phenomena. Whereas, just as there are very few reverse cases of waves converging on a point, there are very few cases more generally where earlier events overdetermine later ones.

> Whatever goes on leaves widespread and varied traces at future times. ... It is plausible that very many simultaneous disjoint combinations of traces of any present fact are determinants thereof; there is no lawful way for the combination to have come about in the absence of the fact.
>
> (Lewis 1979, 50)

Thus the future overdetermines the past to a much greater extent than the past overdetermines the future. That is the asymmetry of overdetermination.

The asymmetry is not total, however. Lewis accepts that the past overdetermines the future occasionally and mildly:

> We have our stock examples — the victim whose heart is simultaneously pierced by two bullets, and the like. But these cases seem uncommon. Moreover, the overdetermination is not very extreme. We have more than one determinant, but still not a very great number.
>
> (Lewis 1979, 49-50)

This is a case of causal overdetermination, which we have not yet discussed. Presumably the thought is that cases of causal overdetermination must also be cases of nomic overdetermination, because causation happens in accordance with the laws of nature.

The asymmetry of overdetermination is probably the most basic and difficult aspect of Lewis's views about temporal asymmetry, and raises many questions. We shall consider some of them shortly (2.4). Before we move on, it bears emphasis that the asymmetry of overdetermination is both a *contingent* and a *basic* fact about the world — the latter less widely recognised than the former, perhaps. The asymmetry thesis is contingent: it concerns the way things are at *this* world. Moreover it is basic, in the sense that it is not reduced to the asymmetry of entropy or to any other known asymmetry — which is the starting point of some objections, as we shall see. In addition, the asymmetry plays a fundamental role in Lewis's theory.

> Let me emphasize, once more, that the asymmetry of overdetermination is a contingent, *de facto* matter. Moreover, it may be a local matter, holding near here but not in remote parts of time and space. If so, then all that rests on it — the asymmetries of miracles, of counterfactual dependence, of causation and openness — may likewise be local and subject to exceptions.
>
> (Lewis 1979, 50–51)

This shows that the asymmetry of overdetermination is fundamental to Lewis's theory: that it is the ultimate explanation of the other asymmetries he mentions, and to some of which we now turn.

2.3.3 The Asymmetry of Miracles

The *asymmetry of miracles* is the fact that small, localised, simple violations of the laws of nature can make a world exactly like ours become unlike ours after the violation, whereas to make a world factually unlike ours become exactly like ours requires big, widespread, diverse violations of the laws of nature. A small miracle is a small, localised, simple violation of law, and a big miracle is a big, widespread, diverse violation of law (Lewis 1979, 47-8). The laws in question are the laws of *our* world (Lewis 1979, 44-5).

The asymmetry of overdetermination gives rise to the asymmetry of miracles:

> I suggest that what makes convergence [of a possible world with ours] take so much more of a miracle than divergence... is an asymmetry of overdetermination...
>
> (Lewis 1979, 49-50)

It works like this. Later facts nomically overdetermine earlier ones, meaning many different sets of later facts nomically determine earlier ones. That means that if a world unlike ours becomes exactly like ours, many subsequent sets of facts nomically (by our laws) determine earlier facts which did not, at the world in question, occur. For every such set of facts, a violation of (our) law must have occurred; and there are many such sets. So violation of the laws entailing these sets will be widespread and diverse. By contrast, earlier facts do not nomically overdetermine later facts to anything like the same degree. This means that a world which is exactly like ours at an earlier time and completely unlike at a later time need not have violated our laws to a great extent.

2.3.4 The Asymmetry of Counterfactual Dependence

The *asymmetry of counterfactual dependence* is the result of the asymmetry of miracles, and the identification of comparative closeness with comparative similarity. We need not revisit Lewis's treatment of the Nixon case — the strategy there is quite clear. Since closeness is similarity, the worlds which make our counterfactuals true must be more similar than other worlds. So we avoid big miracles and look for match of fact, in the way he describes. If we accept his four similarity criteria, we get the asymmetry of counterfactual dependence: for worlds that converge with ours require larger miracles to do so, and are therefore further than some worlds that diverge from ours. So

foretracking counterfactuals stand a chance of truth, whereas backtracking ones do not.

The future similarity objection (the Nixon objection) is, of course, not an objection that concerns backtrackers. It is an objection to the conception of closeness as similarity. But Lewis's answer to it yields the asymmetry of counterfactual dependence, and thus (since it does so deliberately) kills two birds with one stone. I think this is the explanation for a curious feature of Lewis's discussion of the asymmetry of counterfactual dependence, which Bennett reacts to and which we shall now discuss: namely, that a friend of backtrackers in any form surely does not suppose that backtracking counterfactuals entail the world will be just as it is in the future. Suppose I say:

> Mr D'Arcy did not ask for a favour today, but if Mr D'Arcy had asked today, then he and Elizabeth would not have argued yesterday.

In supposing so, surely I do not thereby suppose that Mr D'Arcy's request would have no consequences. Presumably Elizabeth does not actually grant Mr D'Arcy a favour, since he does not actually ask. Does my backtracker therefore commit me to thinking that if Mr D'Arcy had asked today, Elizabeth would not have granted the favour, despite the fact that they would not have argued yesterday? Surely not. But worlds that are dissimilar to ours throughout their whole past and future are more dissimilar than worlds with large areas of perfect match. Thus we are presented with a stark choice between worlds where the future but not the past matches our world completely, and worlds where the past matches but the future does not. Neither seems particularly amenable for interpreting backtrackers.

More generally, even if the asymmetry of miracles does establish an asymmetry in the way that future states of affairs may be greatly affected by small differences in earlier states of affairs, it does not quite establish the asymmetry of counterfactual dependence which Lewis asserts. It is with this point that we shall start our critique.

2.4 Internal Criticisms of Counterfactual Asymmetry

2.4.1 Lewis Needs Backtrackers

There will often be a trade-off between exact match of fact and size of miracle. Suppose Nixon is on the other side of the room, thinking peaceful thoughts, and suddenly — bump — he travels four metres in an instant, and is pressing the button with a scowl. Clearly we envisage nothing of the sort when we suppose Nixon to have pressed the button. To avoid sudden bumps, we sometimes locate the miracle a bit before the antecedent, despite the fact that this loses us some match of actual fact. Lewis prefers a smooth, orderly transition: a neuron somewhere in Nixon's head miraculously fires a few minutes earlier, and subsequently the button is pushed.

Lewis's account must therefore endorse those backtrackers which take us from the antecedent to the small miracle which brings it about. For instance, if Nixon had pressed the button then, miraculously yet inconspicuously, a few neurons would have fired in his brain shortly beforehand. As Bennett puts it:

> [Lewis's theory] provides generously for forward counterfactuals, but the only backward ones for which it makes any provision are the scanty affairs needed to get his closest worlds from likeness to $[w_0]$ to the truth of the antecedent.
>
> (Bennett 2001, 183)

Bennett goes on to point out that the affairs needed to make foretrackers true might not be so scanty after all. Suppose I started to say:

> If dinosaurs roamed the earth today...

It depends on context, but an obvious scenario in which dinosaurs would roam the earth is one in which some meteorite impact millions of years ago was averted, and the dinosaurs persisted. There are obvious interpretations of that antecedent, natural interpretations in many contexts, on which the dinosaurs never died out. In that case, the small miracle would have occurred millions of years ago. So in the context where vagueness is resolved that way, we have backtrackers concerning millions of years. And many of the things we see on Earth today — skyscrapers, the internet — would probably be absent (or at least might be) if the dinosaurs had survived, because we would never

have ventured down from the trees. Of course, in some contexts we *might* suppose dinosaurs to coexist with skyscrapers and the rest — as, for example, in the film *Jurassic Park*. But the technological advances and DNA discoveries featured in that film are not the only, nor even the most obvious, way to suppose dinosaurs roam the earth today. So it is clear that the location of the smallest miracle is highly context dependent: it is not always a matter of backtracking just a little bit to insert an inconspicuous miracle, even if there are opportunities to do so. Sometimes we backtrack a lot: we ignore the possibility that reproductive technologies develop a bit faster and produce live dinosaurs, in favour of a perterbation in the path of an asteroid millions of years ago. Lewis's own account, therefore, licenses backtracking in some more subtle and sophisticated way than it sometimes appears to.

The question about where the smallest miracle might occur suggests a semantics for backtrackers which does not deviate substantially from Lewis's picture. Bennett holds that a respectable semantics must be a "late fork" semantics, like Lewis's: that is, it must assume that counterfactual history matches actual history, and forks off as late as it can. He suggests that true backtrackers are interpreted by worlds where the late fork nevertheless precedes the consequent.

The slack in Lewis's account arising from the trade-off between miracle size and match of fact means that such a theory is conceivable. On Lewis's theory, it remains rather opaque exactly how this trade-off works, and how context affects it, as we have seen. A misfiring of neurons in the darkness of Nixon's skull is a miracle both sufficiently small and sufficiently late to be the obvious candidate for the small miracle leading to the counterfactual button push. In the case of the dinosaurs roaming the Earth today, however, the place we naturally put the miracle is not the latest place we could put it. For some reason, in some contexts we go back a lot further, with a huge cost in match of matters of fact for doubtful gain in miracle-size. The principles governing the appropriate location of miracles, like principles governing so many of our common practices, are unclear, complex and subtle. This raises the prospect that they could, after all, accommodate backtrackers.

Bennett's proposal is not detailed,[12] but it illustrates one way in which

[12]I am referring to his proposal in [Bennett, 2001] and [2003], where he rejects the law-abiding theory advanced in [Bennett, 1984]. I have largely avoided discussion of the relative merits of law-abiding and miraculous theories of counterfactuals. A canonical law-abiding theory of counterfactuals is given in [Jackson, 1987]. Further discussions of the role of laws in theories of counterfactuals include: Lewis 1973b, Goodman 1983, Kvart 1986, Bennett 2003, Pruss 2003.

Lewis's asymmetry of counterfactual dependence might be an overstatement. Perhaps we do hold the past fixed to a much greater extent than the future in counterfactual reasoning. But Lewis pushes this view to an extreme, on which almost every backtracker is normally false. It is this extreme view which plays a crucial role in his theory of causation, as we shall see in Chapters 6 and 7. Bennett's point here, I take it, is that Lewis's own semantics does not fully support this doctrine, because of the slack remaining around miracle size and location. This casts doubt on how sharp a contrast can be drawn between true foretrackers and false backtrackers: every foretracker is surrounded by true backtrackers, and by no means is backtracking always kept to a minimum. I suggest that Bennett's argument gives us an initial reason to doubt Lewis's sharp contrast between false backtrackers and true foretrackers.

2.4.2 Bennett-Worlds

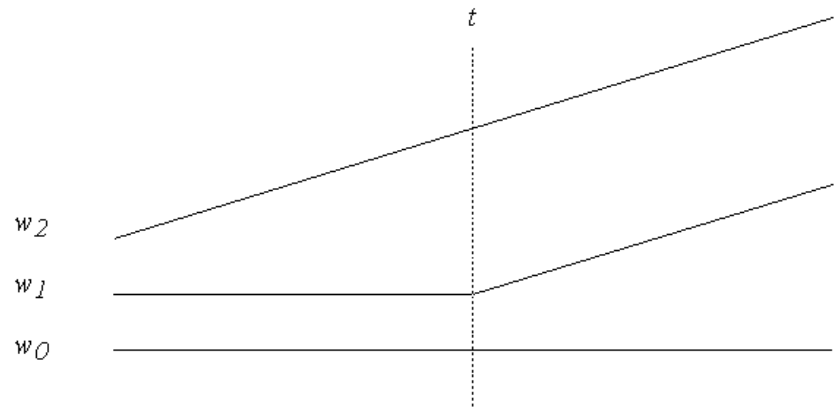
We just considered an attack on the move from asymmetry of miracles to the asymmetry of counterfactual dependence, of the strong flavour which Lewis defends. If this attack succeeds, it is all I need. All I require, for the purposes of my discussion of causation, is some view of counterfactuals on which backtrackers are not automatically false. My discussion of causation could then be seen as an attempt to further delineate the circumstances in which backtrackers are true, and to better understand the more complex asymmetry which Lewis's sharp and simple one parodies. This is indeed how I see the coming discussion. But there are other, perhaps more dramatic weaknesses in Lewis's asymmetry of counterfactual dependence, besides the one we have noted. Understanding some of these weaknesses will provide further motivation for rethinking Lewis's asymmetry thesis.

Now we shall consider an attack on the asymmetry of miracles itself. If the asymmetry of miracles is entailed by the asymmetry of overdetermination, then this is also an attack on the latter.

Bennett asks us to imagine a world w_1 whose history up to time t is just like that at the actual world w_0, and where a small miracle occurs at t.[13] After

[13]It does not matter for present purposes whether t is an instant or a short stretch of time. Let us assume it is an instant, for the sake of simplicity, and let us assume that miracles can be instantaneous violations of law. Like Lewis, I assume that times may be identified across worlds. Lewis says that this assumption may be abandoned "at the cost of some complication" (Lewis 1979, 37), but does not give details. I take it that the complication would stand for present purposes too. For these purposes, one complication that might replace the assumption might be to define t by how much time has elapsed since the start of history, or remains till the end. Or, if there is no start and no end, then for the purposes of

Figure 2.1: Bennett Worlds



t, the world rolls on, obedient to the (deterministic) laws of our world. w_1 is the sort of world which Lewis says makes our counterfactuals true. It is a diverging world, meaning that it shares our history and then diverges by a divergence miracle. By the asymmetry of miracles, divergence miracles may be small.

Bennett points out that the small divergence miracle at w_1 is also an equally small convergence miracle with respect to another world, w_2. To arrive at w_2, simply take a world which matches w_1 after t, remove the small miracle by which we generated w_1 from w_0, and let the past be determined by the laws of w_0. This world, w_2, shares the laws of the actual world strictly, unlike the slightly miraculous w_1. With respect to w_2, the miracle in w_1 is therefore just as minor an infringement of law as it is with respect to our world, w_0. Yet with respect to w_2, the miracle in w_1 is a convergence miracle, and w_1 is a convergent world. For the two worlds share actual fact after t but not before t, and they would share laws if it were not for what happens in w_1 at t.

In Figure 2.1, world histories are represented by lines. Parrallel lines indicate exact match of matters of fact. w_0 is a horizontal line, and w_1 is horizontal until t when it deviates. w_2 is a straight line parrallel to the right-hand portion of w_1.

This argument undermines the asymmetry of miracles as a necessary thesis about all worlds. However, as we saw (2.3.2), Lewis emphasises that it is a contingent thesis, since it rests on the asymmetry of overdetermination, which is contingent. Predictably, then, his response to Bennett's objection is to deny that our world is very much like w_2, even though matters of fact at w_2 converge with those at our world at t. Lewis says:

---

this argument, t can be defined simply as the point before or after which two distinct worlds are exactly alike.

> A Bennett-world is deceptive. After the time of its convergence with w_1, it contains exactly the same apparent traces of its past that w_1 does; and the traces to be found in w_1 are such as to record a past exactly like that of the base world w_0... But the past of the Bennett-world is not like the past of w_0...
>
> (Lewis 1986a, 57)

Lewis suggests that the traces at w_2 would look very much like traces of the past of w_0. But (for all we have a right to think) the past of w_2 may be nothing like the past of w_0. Then the traces would be deceptive, and this, Lewis says, is a big difference between w_2 and our w_0.

Two obvious questions can be asked of Lewis's response. The first is whether deceptiveness of traces can constitute a big difference between worlds. The second is whether Bennett-worlds are indeed deceptive.

Concerning the first question, it does seem that Lewis thinks deceptiveness constitutes a big difference between worlds:

> ...in a world like w_0, one that manifests the ordinary de facto asymmetries, we also have plenty of very incomplete cross sections that postdetermine incomplete cross sections at earlier times. It is these incomplete postdeterminants that are missing from the Bennett-world [w_2].
>
> (Lewis 1986a, 57-8)

But why should it matter if traces are missing at the Bennett-world? Distance between worlds depends on how similar worlds are in matters of general and particular fact. The asymmetry of overdetermination at our world may supervene on matters of general and particular fact, but it is not one itself. Since Bennett-worlds share our laws, by definition of determinism they differ from our world in at least some matters of particular fact at all times. But they might still be fairly close. Not only do they share our laws, but their futures are, in quite a strong sense, possible futures for our world. After all, we arrive at a Bennett-world by first imagining a small miracle at a world like ours, and that is just what we do on Lewis's view when we imagine possible futures of our world. It is not immediately obvious why a world *must* be distant, when it shares our laws and when its latter period is just like ours would have been under some counterfactual supposition.

The more damaging objection, however, arises from the second question: Are Bennett-worlds really deceptive?

Bennett's view of his own worlds seems to have varied. In 1984 he believes that his argument "does not harm Lewis's position" (Bennett 1984, 63). He accepts that the contingency of Lewis's thesis protects it: Bennett-worlds may fail to display asymmetry, but that just shows ours is not a Bennett-world. In 2001, however, Bennett offers a stiffer reply to Lewis's response:

> ...on natural assumptions about the traces left in localized incomplete cross-sections of deterministic worlds, we should say that the localized... traces at w_1 are deceptive, while those at w_0 and w_2 are not... Lewis gives an example to show how a world can be locally deceptive about its past although no miracle has occurred. I concede the general point (though with doubts about the example), but I do not see why a Bennett-world must be like that.
>
> (Bennett 2001, 197–8)

Bennett simply challenges Lewis's unargued claim that a world like w_2 must contain traces of the history shared by w_0 and w_1 up to t. I endorse Bennett's challenge, and offer the following arguments in support.

First, Lewis's proposal is unclear. The notion of a trace is not precise; to do the work Lewis requires, it must be stronger than the strict notion of an incomplete postdeterminant. Surely, almost any fact would qualify as an incomplete postdeterminant, in the sense that more facts could be specified to make a complete postdeterminant. Lewis has something stronger in mind, it seems: not all incomplete postdeterminants are equal. Some facts are more *salient* than others, if we are to pick them out and identify them as traces. But salience is not the only property of traces. Presumably, a given fact — even a salient one — could be combined with a great many other possible facts to entail a great many different and contradictory earlier states of affairs. For a trace to be misleading, as Lewis says Bennett-world traces are, it must be *suggestive*, in the sense that it must somehow suggest certain earlier facts, when it is equally compatible with other earlier facts. Salience and suggestiveness are perfectly intuitive notions, central to inductive inference, but they are not notions that have been captured in the discussion of nomic determination.

Second, in the absence of supplementation by notions such as salience and suggestiveness, there is a fairly obvious reason why w_2 will not exhibit misleading traces. Consider some event[14] e which happens at a time $t-$, before t

[14]The argument does not rely on events: we could use some fact about w_1 before t, or some proposition concerning w_1 before t, true of w_1 but not w_2.

at w_0. Let us suppose e leaves traces such as e', which occurs at time $t+$, a time after t. The rest of the details of the Bennett-world set-up remain; so w_1 is just like w_0 up to t, when a small miracle occurs. Now ask: does the miracle disrupt the lawful process by which the trace e' is nomically determined?

There are two possibilities. The first is that it disrupts the lawful process. Then there is no reason to suppose that e' will be present in w_1. If not, then e' will not be present in w_2 either, since w_2 matches w_1 after t. Thus e' is not a deceptive trace in w_2. And the same goes for every other trace of e at $t+$ depending on that lawful process. None of these traces at w_0 will be deceptive traces at w_2. The other possibility is that the miracle does not affect the lawful process giving rise to e'. In that case, e' will be present in both w_0 and the miraculous w_1. So it will be present in w_2 as well, which matches w_1 after t. But if the laws by which it is nomically determined have not been disrupted, then nor has the nomic determination. Hence e occurs in w_2. Hence e' is not deceptive in w_2, and nor is any other trace of e which is not disrupted by the miracle at time $t+$. What goes for e goes for any other event at w_0 before t. Hence there are no deceptive traces at w_2, in Lewis's sense of deceptive trace.[15]

The latter part of the argument appears secure, but the former relies on the assumption that, in general, a trace e' does not occur unless it is nomically determined by e. Of course Lewis could reject this assumption. Perhaps e' would have come about by a different route, even if e had not occurred. We should certainly accept that this might sometimes be the case. But is there any reason to think this will be so *in general*? Some cases might happen to be like that, but as Lewis himself observes, they are rather rare: they are cases of causal redundancy. If the back-up causes actually occur, then we have nomic overdetermination at the actual world, going from past to future. The idea that overdetermination in this direction occurs on such a massive scale is, as Lewis himself argues, implausible. Moreover it directly contradicts the thesis Lewis defends, that massive overdetermination occurs of past by future but not vice versa. It is only slightly less implausible that there is massive redundancy among unactualised potential causes. And although it might not outright contradict Lewis's asymmetry of overdetermination, it rather goes against the grain to invoke what would amount to a systematic overdetermination of the

[15]Strictly, we should not run the argument for all the traces together, but traces individually. Thus, for any given trace e' of e in w_0, either the lawful process leading to it is disrupted at t or it is not. If it is, then there is no reason to suppose that e' is present at w_1, and hence at w_2, and hence no reason to suppose that e' is misleading at w_2. If the process is not disrupted, e' is present at all of w_0, w_1 and w_2, but not deceptive, for e is also present.

actual future by many unactualised pasts, in order to defend the proposed overdetermination of the actual past by the actual future.

Bennett-worlds share our laws, but have small convergence miracles where we have small divergence miracles. Lewis argues that Bennett-worlds are not much like ours, because they would contain deceptive traces of a past that never happened. Bennett contends that this assertion is unsupported, and I have further argued that it is false. If I am right, then our world could be (in fact, probably is) a Bennett-world — a world to which some world w_1 converges by a small miracle. Supposing so does not involve supposing that our world is radically deceptive, that ordinary inductive inference about the past is insecure, or anything of the sort.

We can conclude that the asymmetry of miracles is to be rejected. If the asymmetry of overdetermination entails the asymmetry of miracles then the former must also be rejected. However the asymmetry of overdetermination may be thought independently plausible, and hence appealed to in defence of the asymmetry of miracles. Or it may, perhaps, be doubted whether the asymmetry of overdetermination entails the asymmetry of miracles. So we shall now consider an argument from Adam Elga to the effect that there is no asymmetry of overdetermination.

2.4.3 Elga-Worlds[16]

Elga produces a highly engaging argument that Lewis's comparative similarity measure does not yield the asymmetry of counterfactual dependence after all, because it does not yield the asymmetry of overdetermination upon which that thesis rests. Elga asks us to imagine a hungry woman called Gretta, intent upon frying an egg for her breakfast. At 7:55am she takes an egg from the refrigerator. At 8:00am she cracks it into a hot frying pan. At 8:05am there is a cooked egg in the pan. According to Lewis, a small miracle at 7:55am could result in Gretta returning the egg directly to the refrigerator (in favour of cornflakes, perhaps). At 8:00am the egg would not be cracked into the pan and at 8:05am there would be no cooked egg. When we say, "If Gretta hadn't cracked an egg into the pan at 8:00am, there would have been no cooked egg there at 8:05am", what we say is true in virtue of the fact that the closest no-crack worlds are of this sort. They are closest in part because the miracle required is small and inconspicuous: a few neurons fire anomalously at the back of Gretta's brain at 7:55am, and she replaces the egg and goes for cornflakes.

[16]I am indebted to Adam Elga for helpful comments on this subsection.

That is Lewis's story. Elga asks us to imagine the egg-cooking process in reverse, as if we were watching a video played backwards. At 8:05am there is a warm cooked egg in the pan. It uncooks over the next five minutes, liquifying and giving up heat to the pan (as well as absorbing some heat from the air). At 8:00am it is raw, and a pressure wave in the air and pan converges on the egg to drive it upwards into its shell. The shell seals, and Gretta walks backwards (rather slowly, presumably) to the refrigerator, turns round, and puts the egg inside at 7:55am.

Elga makes two claims about this sequence of events. The first is that it is possible as far as our *fundamental dynamical laws* go. The only laws it breaks, says Elga, are statistical laws: the laws of thermodynamics. So if we could create a situation just like the situation in Gretta's kitchen at 8:05am, except with the velocities of all the particles reversed, then the next ten minutes would unfold in just the bizarre way we have described. The second claim is that the state at 8:05am is very hard to achieve, because it is very delicate: tiny changes to small groups of particles will disrupt the subsequent course of events so that the egg just sits there, cooling slowly.

> Make a small-miracle change to the end state of a process and run the laws backwards. Certainly the change disrupts the coordinated movement of the process in the neighbourhood of the change. If the parts of the system are strongly coupled, then the region "infected" by the change... will grow rapidly.
>
> (Elga 2000, S322)

And many thermodynamically irreversible processes are indeed strongly coupled, so that change in one part quickly spreads.

Granting these two claims, the consequences for Lewis's various asymmetry theses are as follows. The asymmetry of miracles fails for any strongly coupled thermodynamically irreversible process. We just granted that tiny changes in the velocity-reversed counterpart of Gretta's kitchen at 8:05am would quickly spread, leading to the marvellous backwards cooking processes failing to occur. But this is just to grant that it would only take a small miracle to disrupt the backwards process of uncooking, uncracking the egg, and so on. It is only by the fundamental dynamical laws that the change will spread throughout the system and disrupt it; the process is assumed to be totally deterministic. But it is also totally time-symmetric, since the fundamental laws are reversible. So what goes for the start state of that strange uncooking process goes for the

end-state of the much more familiar cooking process. Therefore it only takes a small miracle in a strongly-coupled, thermodynamically-irreversible process to yield a world with a past quite unlike the actual past, but a future that matches exactly: an Elga-world.

The argument also has consequences for Lewis's asymmetry of overdetermination. Contrary to Lewis's claim, the past overdetermines the future, on exactly the same scale that the future overdetermines the past: many disjoint sets of later facts nomically determine any given earlier fact. But Elga shows that this is false, by showing that small differences in later states of affairs can nomically entail surprisingly extensive earlier consequences. In effect, this shows that there are not as many disjoint sets of later facts as Lewis hoped for. We need to specify an awful lot about the state of Gretta's kitchen before we have a description that nomically entails that she got an egg out of the fridge ten minutes earlier. Among the normal-looking states of Gretta's kitchen, we must rule out many millions of states differing in tiny, humanly imperceptible ways, yet whose backward nomic consequences would be thermodynamically-irregular — that is, utterly weird. It no longer seems as plausible that very many disjoint sets of facts about Gretta's kitchen at 8:05am nomically determine that she got an egg out of the fridge at 7:55am. Thus it no longer seems plausible that nomic overdetermination is asymmetric.

What work, we might ask, is the appeal to thermodynamics doing in Elga's argument? Elga's argument shows that small convergence miracles to the actual world are possible, but I have argued that Bennett's argument already suggests this (though perhaps not so conclusively). What more does Elga's argument achieve?

The appeal to thermodynamics plays for Elga the role that the asymmetry of overdetermination plays for Lewis in his semantic theory of counterfactuals. That is, it secures the spreading-out of consequences of a small miracle. A small change in an end-state of a thermodynamically irreversible process nomically determines earlier states which increasingly differ from the actual world. I take it that the appeal to physics is intended to secure the point.

It is interesting to note that Elga-worlds, unlike Bennett-worlds, are indisputably deceptive. There are many states differing imperceptibly from the state of Gretta's kitchen at 8:05am, but which have come about by a different and weird process. These correspond to the small changes we would have to make to the start-state of the velocity-reversed process, in order to disrupt the subsequent highly delicate uncooking process. These disruptions turn it into

an ordinary cooling and eventual rotting process, which — viewed the right way round again — will become weird unrotting and spontaneous heating processes. Thus:

> It only takes a small miracle to make the difference between the actual world (a world with many veridical traces of the egg-cracking) and w_3 (a world in which those same traces are all highly misleading). In general, the existence of apparent traces of an event (together with the laws, ...) falls far short of entailing that the event occurred.
>
> (Elga 2000, S324)

Thus Elga-worlds are definitely deceptive. But they are also definitely close, by Lewis's measure, since they are arrived at by imagining our world time-reversed, inserting a single small miracle with spreading consequences, and then righting the temporal order again. The miracle remains small because the laws it violates are time-symmetric.

Previously I argued that any trace e' would be present in a Bennett-world if and, normally, only if e itself was present. Elga's argument seems to show that close worlds can have deceptive traces of a past which, at those worlds, didn't happen. Is there a tension? No, because Elga is not claiming — as Lewis was — that traces are deceptive at lawful worlds to which other worlds converge by a small miracle. The problem with that claim was the role of the miracle: either it affected the traces, in which case removing it would remove the traces, or it did not affect the traces, in which case the traces were presumably as veridical as in any other world. But Elga-worlds are deceptive for a quite different reason: because they amount to very unusual starting situations for processes in which the statistical laws of thermodynamics are violated. Miracles have nothing to do with the deceptiveness of an Elga-world. The state of Gretta's kitchen at 8am, when the raw egg is in the pan, is deceptive in the Elga-world we have discussed, even though the miracle does not occur until 8:05am. For although it looks just like the actual egg, which has been cracked into the pan, in the Elga-world the egg has formed by a process of unrotting. Elga-worlds are deceptive for an entirely different reason: because they have freakish and improbable pasts but become very, indeed humanly indiscernibly, similar to worlds that are thermodynamically standard. No miracle is require for them to do this; the miracle is required only to secure *exact* subsequent convergence. We could do away with the miracle, to produce a Bennett-Elga-world: a world

with a freakish but fundamentally law-abiding past, and a future which does not exactly match ours but which is exactly as likely as ours to unfold in some humdrum, thermodynamically normal way. Even though this would not be how our world unfolds, the Bennett-Elga-world would still be highly deceptive about its past.[17]

2.5 Summary

We have rehearsed the elements of Lewis's semantics for counterfactuals, and the associated proposal that counterfactual dependence is radically asymmetric — to the extent that backtrackers are, except in special circumstances, false. And we have endorsed a selection of criticisms of this view. The asymmetry cannot be a simple matter, since every foretracker whose antecedent is not itself the smallest miracle that makes it true will semantically entail backtrackers concerning the interval between itself and that little miracle. As we have seen, even when we are assessing foretrackers, we sometimes backtrack more than Lewis's presentation of the asymmetry of miracles would suggest that we strictly have to. If we sometimes backtrack more than Lewis's presentation suggests when assessing foretrackers, this raises the prospect that we might, without deviating any further from Lewis's principles than we already do when assessing foretrackers, sometimes backtrack when assessing a backtracker. So perhaps backtrackers can be accommodated on a Lewis-style theory after all.

We have also endorsed two more robust criticisms. Bennett-worlds undermine the asymmetry of miracles and Elga-worlds undermine the asymmetry of overdetermination. Although these criticisms are more robust, it is important to see that they are, in a way, less significant. In the end, they are merely internal criticisms of Lewis's efforts to capture an alleged asymmetry. They show that Lewis's theory does not yield an asymmetry, and thus that it does not succeed in capturing one that really exists. It is important that we rehearse these attacks, but for our purposes, it is more important that we deny there is an asymmetry of the sort Lewis is seeking to capture. This is the purpose of the next chapter. In this chapter, I sought to rebut the efforts of Lewis to capture the asymmetry of counterfactual depedence. In the next, I seek to rebut the reasons for thinking there is a sharp temporal asymmetry among counterfactuals.

[17] An excellent source of further discussion of the philosophical implications of thermodynamic asymmetries is Huw Price (eg. Price, 1992a,b, 1996, Menzies and Price, 1993).

Chapter 3

Backtrackers Can Be True

3.0 Abstract

The purpose of this chapter is to argue that backtrackers can be true, and to provide a method for working out whether they are true. I argue that backtrackers and foretrackers are not confined to distinct contexts, as Lewis suggests: they are compatible, and can be true in the same context. I argue further that the grammatical complexity with which we prefer to express backtracking reasoning need not indicate a relevant shift of context. Whether an asymmetry of counterfactual dependence exists and what its nature might be is, therefore, an open question, not to be settled by appeal to grammar. Against that background, I suggest that we need a method to help us tell when backtrackers are true, irrespective of distracting linguistic features. This amounts to a further criticism of Lewis's asymmetry of counterfactual dependence, because it deprives that thesis of its explanandum — an ordinary context in which the resolution of counterfactual vagueness makes backtrackers mostly false. Lewis's "visualisation" method is criticised, and the Ramsey Test is found unsuitable as it stands. Drawing on hints from Frank Ramsey, Nelson Goodman and Dorothy Edgington, I propose a simple test which seeks to harness our robust and temporally neutral abilities to make and assess inductive inferences for the task of assessing counterfactuals. I dub it the *Inference Test*.

3.1 The Compatibility of Backtrackers and Foretrackers

Take an event, and then suppose it away, first asking about the later consequences, and then asking about the earlier events leading up it. There are occasions when, if you suppose both away at once, you seem to risk saying strange things. I am holding a glass, but suppose I had dropped it. We might naturally suppose that, if the glass hadn't broken under those circumstances, I would have been rather surprised. Now suppose we assert that, if the glass hadn't broken, I wouldn't have dropped it. There is a strong intuition that these two counterfactuals are in tension: if the glass hadn't broken, either I would have been surprised at the non-breaking glass, or I would not have dropped the glass, but not both. For then why would I have been surprised?

Contrary pairs of counterfactuals have the form $A > C$ and $A >\!\sim C$. Here, we have counterfactuals which are not direct contraries, being of the form $B > C$ and $B > A$; yet it seems they are not true together. Suppose that at A-worlds I do not drop the glass, at B-worlds the glass does not break, and at C-worlds I am surprised. If the closest B-worlds are also C-worlds, then it seems they are not A-worlds: for C-worlds are worlds where I am surprised, and that surprise would itself be surprising if I didn't drop the glass. Likewise, if we suppose the closest B-worlds are A-worlds, it seems they will not be C-worlds: once again, worlds where I do not drop the glass give me no cause for surprise at the unbroken glass.

Lewis explains this phenomenon by suggesting that counterfactuals are vague. It is possible to know the facts you would need to decide whether a counterfactual is true, even to be an expert in the relevant field, yet remain undecided. For example, even an accomplished historian may be uncertain about this:

(A) If Caesar had been in charge in Korea, he would have used nuclear weapons.

The historian's uncertainty may be brought out by asking her to consider another counterfactual.

(B) If Caesar had been in charge in Korea, he would have used catapults.

The expert knows enough about Caesar's competence as a military commander to be confident that he would not have used both catapults and nuclear

weapons in Korea. So she regards these as contraries, in the simple sense that they are not true together.[1]

On Lewis's view (Lewis 1973b, 1979), the explanation is that counterfactuals are vague, and until some of this vagueness is resolved, they lack determinate truth-value. Vagueness is resolved by context: once it is understood that we are discussing military psychology with reference to the Korean War, it is clear that (A) expresses an interesting hypothesis, and (B) is facetious.

This point about counterfactual vagueness is a familiar one, and shows that we should be cautious when mixing counterfactuals generally that we are not illicitly switching between different resolutions of vagueness. Resolving the vagueness of one counterfactual may render another counterfactual false under that resolution.

On its own, this point shows nothing special about backtrackers. Nevertheless, Lewis suggests that backtrackers and foretrackers are not true under the same resolution of vagueness. Examples like the one we began with (the breaking glass) show that admitting backtrackers and foretrackers together gives rise to all sorts of trouble. This is an idea Lewis gets from Downing and Bennett (although the latter has subsequently revised his opinion). We can illustrate with a story.

> Elizabeth and Mr D'Arcy argued yesterday, so if Mr D'Arcy asked Elizabeth for a favour today, she would not grant it. But Mr D'Arcy is proud, and if they had argued yesterday, he would not have asked Elizabeth for a favour today. This means that, if Mr D'Arcy had asked Elizabeth for a favour today, they wouldn't have argued yesterday. But normally Elizabeth is kind: thus if they hadn't argued yesterday, she would happily grant Mr D'Arcy a favour. So if Mr D'Arcy asked Elizabeth for a favour today, she would grant it.[2]

We have contradicted something we previously found plausible, by arriving at its contrary; and our reasoning employed backtrackers.

The moral drawn by Downing, early Bennett and Lewis is that backtrackers and foretrackers do not mix. Lewis further explains that we ordinarily resolve the vagueness of counterfactuals in a way that favours foretracking. In order to give speakers their best chance of saying true things, we can be persuaded to

[1] This much-cited example originates at (Quine 1960, 222).

[2] Lewis replaces Elizabeth and Mr D'Arcy with Jack and Jim (Lewis 1979, 33–4).

backtrack, charitable conversants that we are; but we slip naturally back into a foretracking resolution as soon as the special backtracker-friendly contexts have passed.

Indeed it does seem that we are persuaded to resolve counterfactual vagueness differently in this story, as we were in the story about the breaking glass. This fact completely fails, however, to support the contention that foretracking and backtracking resolutions of vagueness are incompatible. It is this contention, I take it, which underlies the claim that the ordinary resolution is (largely) a foretracking one, and thus which the asymmetry of counterfactual dependence is directed at explaining. But, I shall now argue, the incompatibility of backtrackers and foretrackers under the same resolution of vagueness is not demonstrated by this line of argument. In fact, there is no problem mixing backtrackers and foretrackers; hence this is no reason to suppose they cannot both be true under ordinary resolutions of vagueness.

The mere fact that backtrackers and foretrackers are *sometimes* incompatible does not entail any such strong conclusion. After all, the example of Caesar in Korea shows that foretrackers are also *sometimes* incompatible in just the same way, and clearly it is possible to hold more than one foretracker true under a given resolution of vagueness. A later Bennett reconsiders the argument about Elizabeth and Mr D'Arcy, and makes the following diagnosis of the problem. First we are asked in counterfactualising to attend to Elizabeth's anger and ignore D'Arcy's pride, yielding the first claim, "If D'Arcy asked Elizabeth for a favour today, she would not grant it". Then we are asked to attend to D'Arcy's pride and to ignore Elizabeth's anger, says Bennett, yielding the counterfactual "If D'Arcy were to ask Elizabeth for a favour she would grant it". The backtracker in the middle is not the source of the trouble:

> The [Elizabeth and Mr D'Arcy] example has nothing essentially to do with temporal direction, so it tells us nothing about backward counterfactual conditionals...
>
> (Bennett 2001, 180)

This is just the diagnosis I used to generate the example of the breaking glass with which we started. But as Bennett points out, it has nothing to do with backtracking.

We can illustrate the weakness of the Elizabeth and D'Arcy argument by considering the following story:

> Elizabeth is kind, and well-disposed towards Mr D'Arcy. So if Elizabeth and Mr D'Arcy hadn't argued yesterday, she would happily grant Mr D'Arcy a favour. So if Mr D'Arcy asked Elizabeth for a favour today, she would grant it. But wait: Elizabeth and Mr D'Arcy have some issues to work through. If Elizabeth and D'Arcy had not argued yesterday, they would have argued today. In that case, if Mr D'Arcy asked Elizabeth for a favour today, she still wouldn't grant it.

The addition of underlying issues in this example plays the role of Mr D'Arcy's pride in the previous example. It induces us to resolve the vagueness differently, holding different facts fixed under our supposition, and we end up with an apparent contradiction. But all the counterfactuals here are foretrackers: they are mostly recycled from the previous story, and the only one we added is clearly a foretracker (if they had not argued yesterday, they would have argued today). If this line of argument shows that backtrackers and foretrackers cannot be true under the same context, then by parity of reasoning, it shows that foretrackers cannot be true under the same context. But that is absurd.

It is clear that, when we resolve the vagueness of a counterfactual, some other counterfactuals may be false under that resolution. It is further clear that some backtrackers might come out false when we resolve vagueness favourably for some foretracker, and vice versa. This does not show that we can never hold a backtracker with a foretracker, any more than the debate about Caesar's choice of weoponry in Korea shows we can never hold two foretrackers true. The stories which are told to support the claim that backtrackers and foretrackers mix badly can be matched by stories which, by parity, would licence the absurd conclusion that foretrackers mix badly. In fact, all such stories show is that context is sensitive and easily manipulated, and that care must be taken when mixing counterfactuals of *any* temporal direction. This removes one reason for denying that backtrackers and foretrackers may be true under the same resolution of vagueness.

3.2 Grammar

The other reason that Lewis takes backtrackers to be awkward is that their *expression* is often awkward.

> Back-tracking counterfactuals, used in a context which favours their truth, are marked by a syntactic peculiarity. They are the

> ones in which the usual subjunctive conditional constructions are readily replaced by more complicated constructions: "If it were that ... then it would have to be that ..." or the like.
>
> (Lewis 1979, 34–5)

This, too, is supposed to be explained by the thesis of contextual resolution of vagueness: the grammatical (syntactic) complexity is a way of adjusting context to favour a special, backtracking resolution of vagueness. However, we shall see that not all backtrackers are accompanied by grammatical complexity, even if most are. The existence of a handful of backtrackers that, like most foretrackers, are simple grammatical subjunctives does not tell decisively against Lewis's view, because any generalisation about a practice as complex as natural language is highly likely to have a few exceptions. But it is odd nonetheless, because on Lewis's view, these backtrackers are actually false under a standard resolution of vagueness, despite the fact we assert them without any of the contextual markers of a "special" backtracker-friendly resolution of vagueness. If such a resolution is occurring, its mechanism is mysterious; there are no signs, apart from the truth of the backtrackers in question. If it is not occurring, then these backtrackers are strictly false. Further, many backtrackers which we have an aversion to expressing with a simple grammatical subjunctive do not seem to be false so much as odd. Lewis's explanation for their oddity is falsity — which is not the same thing, and may further disrespect our intuitions. But the avoidance of a falsifying resolution of vagueness is not the only possible explanation for the grammatical complication of backtrackers. If the grammatical subjunctive is often taken to have other implications, causal ones for instance, then, I shall suggest, it becomes quite obvious why we would not assert backtrackers using that construction.

Some backtrackers are intuitively true. For example:

Holmes' Hypothesis. Sherlock Holmes finds the fingerprints of the Vice-Provost on the door of the Library of King's College. He hypothesises that if the Vice-Provost hadn't entered the Library, he wouldn't have left his fingerprints on the door. From this hypothesis, and the prints, Holmes deduces (by modus tollens) that the Vice Provost entered the Library.

Holmes' Hypothesis is contingent, of course. The Vice-Provost could have merely touched the door, but not gone in. Nevertheless, it is a perfectly intel-

ligible hypothesis, and quite a reasonable one. We can, without special discomfort, entertain the supposition that it is true. But it is evidently a backtracker, because the prints are left before the Vice-Provost enters the Library. Moreover it is not marked by any of the markers of special conversational context, in the way we should expect if the vagueness resolution story about backtrackers were correct. Grammatically, it is a straightforward subjunctive.

The importance of Holmes' Hypothesis is not merely that it is intuitively obvious. It is also used in a piece of reasoning, an instance of counterfactual modus tollens. It is one thing to deny a backtracker that has been dreamed up for the purpose of argument. It is another to deny a backtracker which has been used in earnest. A legal case which featured Holmes' Hypothesis would not be thrown out, merely on the basis that Holmes' Hypothesis is a backtracker. More substantive grounds would need to be adduced, such as the possibility that the Vice Provost touched the door but did not go in. But if that and similar possibilities are ruled out, there seems to be no further, principled reason why Holmes' Hypothesis might not be true. It certainly seems unlikely that any philosophical argument against backtrackers would defeat Holmes' case in court.

Here is another example of an intuitively acceptable backtracker.

Steaming Coffee Counterfactual. I have a steaming cup full of hot coffee before me, but if it wasn't steaming, it wouldn't be full of hot coffee.

It might be objected that this is not a backtracker, because the heat and the steaming are simultaneous. I have no objection to such a reading, but Lewis does. For his theory of causation must identify the above counterfactual as a backtracker if it is to rule out the steam as a cause of the heat of the coffee (Lewis 1973a, 170–1). This counterfactual is a direct counterexample to Lewis's solution to the problem of effects, since it is obvious that the hot coffee causes the steam and not vice versa.

The Steaming Coffee Counterfactual seems quite intuitive to me. Certainly to deny it seems intuitively difficult. We would have to assert that, if the cup were not steaming, it might nevertheless be full of hot coffee: perhaps a tight lid could have been fitted. But this is a china mug with no lid: surely we can rule that world out as more distant than some where it contains no hot coffee. And the lid is the least outlandish way I can think of to stop the coffee steaming, while leaving the coffee hot. High atmospheric pressure and high temperature, miraculous happenings just above the water, and so on are all

rather far-fetched. There are, surely, closer worlds where the cup is not full of hot coffee — it is empty, or nearly so, or else the coffee has gone cold, having been poured twenty minutes earlier. And the thought that there must be closer worlds like these is borne out very clearly when we ask what world we would think we were in, faced with a cup which was not steaming.

Exhibition of intuitively true backtrackers is suggestive of some problem with Lewis's diagnosis of the problem with backtrackers. According to Lewis, backtrackers are true in special backtracking contexts only; if these are not carefully set up, the backtrackers sound bizarre, because they are false in the default foretracking context (Lewis 1979, 34). Why, then, does there appear to be a handful of exceptions? Lewis's proposal amounts to an error theory about certain, rather unpredictable corners of our common discourse and reasoning processes. Holmes' Hypothesis would not be thrown out of court, so it seems that "common" may include professions such as law or medicine, where these reasoning processes really matter. Likewise scientists and historians might occasionally lapse unwittingly into falsity; and perhaps even — who knows? — careless philosophers. The exceptions are rare, but by their nature we do not notice them when they occur.

On its own, a handful of exceptions will not be conclusive; natural language is a complex thing, and a generalisation may be true enough despite them. But there is another, related problem. Consider this backtracker:

Odd Caesar Counterfactual. If Caesar had used the atom bomb in Korea, he would have been in charge in Korea.

It is hard to make sense of the Odd Caesar Counterfactual without a bit of thought. Yet it is, presumably, fairly plausible: if Caesar had used the atom bomb in Korea, then he would have to have been in charge there, in order to get his hands on an atom bomb and deploy it there. But this argument, even if it is accepted, does not remove the strangeness. It simply is not something we would say. This seems to undercut Lewis's explanation of strangeness in terms of its falsity in a certain context. The Odd Caesar Counterfactual remains strange even in a context where we have convinced ourselves it is true; a fortiori, it remains strange even when its vagueness is eliminated.

Note, too, how different the Odd Caesar Counterfactual is to (A) and (B), which we considered earlier. There, clarification made it clear whether Caesar would use nuclear weaponry or catapults. Neither (A) nor (B) was odd, and they both remained perfectly non-odd even when vagueness was resolved to make one true and the other false. The false one was no odder than the true.

Previously I argued that declaring intuitively *acceptable* backtracking subjunctives false (such as Holmes' Hypothesis and the Steaming Coffee Counterfactual) amounts to an error theory about certain corners of our discourse. The present problem is the complement of the previous one. It concerns declaring intuitively *odd* backtracking subjunctives false. The problem is that falsity fails to remove oddness. The point may be brought out, as I have tried to do, by exhibiting intuitively unacceptable backtracking subjunctives which, after a bit of thought and persuasion (even, if you like, context-shifting and vagueness resolution) we think are true. The point is that they still sound odd. We would prefer to phrase them with a more complex construction even if we accept they could be understood as expressing something true. Therefore the purpose of the more complex expression cannot be to resolve vagueness in such a way as to avoid falsity.

Falsity in context due to an unfavourable resolution of vagueness is not the only possible explanation for the oddity of many backtracking subjunctive conditionals. That might be explained if the grammatical subjunctive has further conversational implications (just as "if" often implies "only if"). Notably, the subjunctive sometimes suggests a causal connection; if so, then speakers would naturally be wary of using it when this implication might create misunderstanding. Holmes' Hypothesis and the Steaming Coffee Counterfactual might be exceptions to this conversational implication, perhaps because it is so obvious that steaming does not cause coffee to be hot, and that fingerprints do not cause entry. We need not develop any such theory; we need merely note that other explanations for the oddity of backtracking subjunctives might be developed. An obvious starting point would be to examine the psychological connection between subjunctives and causation. But this is a task for psychologists, and one which they take seriously (cf. McEleney and Byrne 2006).

We should be wary of taking surface grammar as a logical or philosophical guide. To do so can generate apparent problems where there are none. I have suggested that one such merely apparent problem is the grammatical complexity associated with backtracking counterfactuals, and the corresponding oddity of backtracking grammatical subjunctives. Grammatical complexity arises, not from any deep conceptual feature of counterfactuals, but from a concern not to be misunderstood. The scope for misunderstanding might arise, I have suggested, partly from the fact that the subjunctive construction is often used to express or explain causal connection of some sort. But making this suggestion precise — precisely specifying under which situation a subjunctive construc-

tion suggests what, if any, causal connection — is a linguistic or psychological task, not a philosophical one.

3.3 Counterfactual Asymmetry: An Open Question

Backtracking counterfactuals are often thought to be awkward. Lewis identifies two kinds of awkwardness: the possibility that mixing backtracking and foretracking reasoning will lead to contradictions, and the fact that many backtracking counterfactuals are not naturally expressed with simple grammatical subjunctives, but with some more complex construction. Lewis's solution is to attribute the awkwardness of backtrackers to special features of the way the ubiquitous vagueness of counterfactuals is resolved. According to Lewis, when we resolve vagueness of counterfactuals, we usually do so in a way which rules out backtrackers as false; the exceptions are those occasions when we allow backtrackers and rule out foretrackers.

Nothing I have said so far shows this picture to be untrue. Perhaps we do indeed resolve the vagueness of counterfactuals in favour of one temporal direction at a time. But I have sought to remove the reasons Lewis gave to think that we do this. Lewis argues that mixing backtrackers and foretrackers leads to contradictions, because it involves flipping between different resolutions of vagueness. I pointed out that merely exhibiting an example where such flipping indeed occurs does not show that it must always occur when we mix backtrackers and foretrackers. Lewis also argues that the awkward sound of backtracking subjunctives can be explained by their falsity in ordinary contexts. I argued that this was no explanation because foretrackers do not sound odd even when we decide they are false in context. Lewis also suggests that the complexity which often accompanies backtracking expressions is an indication that context is being manipulated to favour a different, backtracker-friendly resolution of vagueness. I have argued that these indicators are not always present: there are some intuitively true backtrackers.

In short, I have denied that the resolution of vagueness has any special application to the question of how backtrackers and foretrackers relate. One implication of these arguments is that surface grammar is not a reliable guide to logic. Implicit in my discussion is the view that counterfactuals are not to be identified with natural language sentences of a given form. What, then, are counterfactuals? An obvious answer is that they are propositions satisfying an

appropriate semantics. Another option would be to identify counterfactuals with some sentence type in an artificial language employing "$>$". We do not need to decide; all the foregoing discussion implies is that counterfactuals are not to be identified with a natural language sentence type. Any view which allows various grammatical constructions to express counterfactuals (and that includes Lewis's) must identify counterfactuals with something other than a single natural language sentence type. This puts "$>$" on a par with many truth-functional sentential connectives such as "$\vee$" and "$\supset$", both of which differ somewhat from their closest English counterparts.

The vagueness resolution story implies that backtrackers are usually false, except in a few special cases; and then, foretrackers are false. Rebutting this view thus removes a perceived obstacle to appealing to backtrackers in philosophical theories, such as the analysis of causation which I am about to propose. It also removes a central reason to accept Lewis's asymmetry of counterfactual dependence, since this was supposed to provide a semantics for a standard foretracking resolution of counterfactual vagueness. I have argued that, whether backtrackers are standardly false or not, resolution of vagueness to make them false would not address either of the two difficulties which it has often been thought to address. Whether backtrackers are true, how often and in what circumstances, is an open philosophical question: and not, I have argued, one that should be closed by appeal to grammatical features of natural language.

If it is an open question what asymmetries of counterfactual dependence exist, two questions suggest themselves. The first is: are there any good positive reasons to think backtrackers are actually true? The second is: if so, which backtrackers are true? The next three chapters can be seen as giving partial answers to both questions. But before we can address either, we need a tool: a method for assessing the truth of counterfactuals without being prejudiced by their temporal direction.

3.4 Assessing Counterfactuals

3.4.1 Lewis's Visualisation Method

Lewis's semantics comes with an unofficial heuristic for assessing counterfactuals. It goes like this. Imagine a world just like ours, and insert the smallest, most inconspicuous change you can think of in order to make the antecedent come true. Sometimes the smallest change might just be the antecedent com-

ing true, but often it will do less violence to the laws of nature to backtrack a little and insert something inconspicuous earlier. For example, rather than imagining Nixon instantaneously teleporting from his leather chair across the room to the control panel and pressing the big red button, you might imagine a few neurons firing anomalously in his head a few seconds earlier, leading him to get up, walk over to the button and press it. Then imagine the world unrolling according to the laws of our world, and see whether the consequent of the counterfactual comes true. If so, the counterfactual is true; otherwise, it is false.

I call this a *visualisation method*, to emphasise the fact that it relies heavily on our imaginative faculties. This is problematic in several ways. Imagination does not readily respect the laws of nature in the way that the method asks it to. It is quite hard to know what the smallest violation of the laws of nature will be. It is also difficult to tell how affairs would have progressed according to the laws of nature after that small miracle. Moreover even if we think we can guess, there is no reason to suppose that what seems to us like the smallest miracle would really be the smallest miracle, nor that what seems to us like a lawful progression of events thereafter would really be lawful. In short, our imagination is no guide to the comparative similarity of possible worlds, because it is not connected in a truth-tracking way to the determinants of comparative similarity. There is no reason to suppose we will correctly imagine how a world will unfold according to the laws, nor will we notice if we make a mistake. Likewise there is no reason to suppose we can correctly compare the magnitude of various factual differences from our own world, nor any reason to suppose we will notice if the one we choose is not the smallest. The visualisation method is hardly a method at all: it is more like a structured guess.

It might be wondered whether Lewis really means to propose a method that involves imagination in such a central way. Perhaps not; he never explicitly proposes any method for assessing counterfactuals. This heuristic is just the application of the imagination to his semantics for counterfactuals; but if indeed it is used, it is nowhere spelled out explicitly. It may look a bit silly when it is spelled out, but that is not due to a misrepresentation. Indeed a surprising number of arguments and counterexamples appear to be decided in this way: by leaning back, squinting into the middle distance, and visualising worlds. When the method is stated explicitly, it suffers from an obvious weakness, employing a faculty (imagination) which is not particularly well suited to

discovering the truth about the subject to which the method is applied (laws of nature and comparative magnitude of factual differences).

The Lewisian heuristic is supposed to yield the truth about counterfactuals because it involves thinking about the things that actually make counterfactuals true: comparative similarity relations between possible worlds. This gives rise to another criticism of the method. Even if somebody does believe that they have the cognitive faculties for applying the visualisation method — either due to a high opinion of their faculties, or because some cases are thought to be very clear and simple — the method is thoroughly theory-laden. If the results of thought experiments are considered to be a bit like the data produced by ordinary scientific experiments, then it is important that our data are as free from bias as they can be. Otherwise the data might wrongly protect a theory from falsification, or might wrongly falsify a worthy competitor; and if there are competitors, the data may not be agreed upon by proponents of other theories. The visualisation method is extremely biased towards Lewis's semantics. In the first place, we are asked to imagine the smallest change before the antecedent, and then roll the world forwards. This automatically renders the method incapable of showing any backtracker true unless it is one of the handful concerning the interval (if there is one) between the antecedent and the preceding small miracle. Moreover, the method cannot readily be reversed. It is very hard to imagine a world just like ours, imagine the smallest miracle after the antecedent which would lawfully entail the antecedent, then roll the world backwards according to the laws and see whether the consequent comes true. The fact this is harder might be due to the falsity of backtrackers, of course. But then again, it might not. We experience the world in a particular temporal direction, and our cognitive limitations might reflect that, rather than any truth about the direction of counterfactual dependence. Moreover, even if we could reverse Lewis's heuristic, it would not yield a very fruitful method. Clearly, even a proponent of backtracking counterfactuals will not want to endorse a method which yields the result that if the past were different, the future would be exactly as it actually will be.

The visualisation method is therefore unsatisfactory for two reasons: it is of doubtful reliability; and it is informed by theory in a way which produces bias.

Timothy Williamson, in an unpublished paper [Williamson, 2007], has given more attention to the prospects for visualisation as a method of assessing counterfactuals. His suggestion is that we use our usual expectation-forming

capacities to arrive at knowledge of counterfactuals. This is a very interesting contribution to the epistemology of counterfactuals, and it does not succumb to the criticism that we cannot properly imagine worlds unfolding according to the laws. On Williamson's view, it is our ordinary expectation-forming capacities which are applied to counterfactual scenarios: these have a solid basis in everyday activities. The view I adopt in 3.4.3 below has a lot in common with this line of thinking. Note, however, that an expectation-based epistemology of counterfactuals still seems to struggle to explain our knowledge of backtrackers, at least until explicit attention is given to the question of how expectations can be formed about the past. Expectations are temporally asymmetric, so an expectation-based theory will display temporal bias.

3.4.2 The Ramsey Test

Another candidate for an epistemic method for assessing counterfactuals arises from a famous footnote of Ramsey's:

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q...
>
> (Ramsey 1978, 145 — footnote 1)

On this slender basis, a considerable literature has grown up around the Ramsey Test for acceptability of a counterfactual. According to this test, in order to decide whether a counterfactual is true, you add the antecedent to your "stock of beliefs", and then decide on that basis whether the consequent is true.

The Ramsey Test does not suffer from the same theory-ladenness as Lewis's test, but it does suffer from some difficulties. Most centrally, it will not do to hold fixed every belief logically independent of the antecedent. For that will, in many cases, include the consequent. Therefore some heuristic of belief revision is called for. It is widely supposed that we need an account of minimal belief revision: a set of principles governing how we should adjust other beliefs when we entertain a supposition, on the assumption that we should avoid gratuitous adjustments, and yet make more adjustments than logic requires. "Minimal" may be a misleading term, since really what we want is something more like accurate belief revision. We want a method of adjusting beliefs in light of suppositions which reflects how the things those beliefs are about would

change if the suppositions were true. Lewis's heuristic at least has this going for it — that its recommendations concerning belief revision are motivated by a view of how things would be different, were the suppositions in question true.

So the Ramsey Test, as it stands, will not help us. It requires supplementation with a heuristic for belief revision before it can provide determinate pronouncements on given counterfactuals. This point can be brought out sharply by noticing that the results of the Ramsey Test depend on *how* we add P to our stock of beliefs. As Richard Bradley puts it:

> We might suppose that as a matter of fact P is true; in which case we would revise much in the way that we do when we learn of P's truth. Minimal revision in this case might require us not to give up any firm beliefs not contradicted by P. Alternatively, we might suppose or imagine that, contrary to the facts, P is true. A supposition of this kind may be best accommodated by giving up some of one's beliefs not contradicted by P, to allow retention of well-entrenched ideas about the way the world works.
>
> (Bradley 2007, 3)

Thus, as it stands, the Ramsey Test is ambiguous between material and counterfactual conditionals. We cannot use it, therefore, to provide a simple test for counterfactual truth: the directive "Add P to your stock of beliefs, then see whether you believe that Q" is ambiguous, depending on how P is added. This is clearly a difficulty, but to respond to this problem would, surely, be unnecessarily ambitious in the present, rather pragmatic, context. All we want is a rough-and-ready method for helping us assess counterfactuals. We do not need a fully-fledged theory of the psychology of that assessment, or of the principles governing it. So if we can, we should try to avoid answering all the problems thrown up by the Ramsey Test. To do this, I suggest we pull back from the more ambitious reading of the test, and focus more generally on the connection between inference and counterfactuals.

3.4.3 The Inference Test

Ramsey says:

> 'If p, then q' can in no sense be true unless the material implication $p \supset q$ is true; but it generally means that $p \supset q$ is not only true but deducible or discoverable in some way not explicitly stated. This

> is always evident when 'If p then q'... is thought worth stating even when it is already known either that p is false or that q is true. In general we can say with Mill that 'If p then q' means that q is inferrible from p... together with certain facts and laws not explicitly stated but in some way indicated by context.
>
> (Ramsey 1978, 144 — my symbolism)

The mention of inferribility is potentially very useful here. The important point from our present, pragmatic point of view is the suggestion that we see the move from antecedent to consequent as an *inference*. Inferences are things we have a good practical grip on, therefore they are good test material.

Goodman picks up a similar line of thinking:

> When we say
>
> > If that match had been scratched, it would have lighted,
>
> we mean that conditions are such — ie. the match is well made, is dry enough, oxygen enough is present, etc. - that "That match lights" can be inferred from "That match is scratched."
>
> (Goodman 1983, 8)

Goodman develops an account of counterfactuals from this starting point. But the question of whether inferences *tell* us something about the truth of counterfactuals may be distinguished from the question of whether counterfactuals are to be *identified* with inferences in some form. I shall endorse a positive answer to the former question, but remain agnostic on the latter.

Dorothy Edgington displays sensitivity to some of the issues we are touching upon. She suggests we ask what counterfactuals are for.

> Why do we evaluate counterfactuals the way we do? What would go wrong for us if we chose to evaluate them according to the 'standard picture'? The question deserves more attention than it has had in the vast literature on counterfactuals.
>
> (Edgington 2004, 23)

Edgington suggests that counterfactuals are central in "empirical inferences to conclusions about what is actually the case". Since inferences are things which

we make and assess all the time, they might provide the basis for an epistemic method for assessing counterfactuals. Let us explore the possibility further.

The suggestion, then, is that there is a connection between our belief that $A > C$ and our willingness to assent to an inference from A to C. The task we now face is to convert this remark into something like a test, which can be applied to counterfactuals to discover their truth. I propose to adopt the following test for counterfactual truth.

The Inference Test. $A > C$ iff C may be inferred from A.

The phrase "may be inferred" requires elucidation. It does not mean, for example, that if it were the case that A, then some agent would infer that C. For if it were the case that A, the agent may be absent, deluded, or otherwise engaged. Nor does it mean that, if it were the case that A, then some actual agent infers that it would be the case that C. This does not make sense: how can an actual inference be counterfactually conditional upon an unactualised possibility? The right reading lies somewhere between these two wrong suggestions. The phrase "may be inferred" should be given a normative reading, synonymous with "it would be acceptable to infer". The phrase thus combines the two readings recently rejected. The thought is that $A > C$ iff, if A were the case, then an inference to C would be acceptable, by our actual standards.

Of course, this does not eliminate dispositions: acceptability is a disposition, even qualified with "actual". The presence of unanalysed dispositions may be thought troubling, because dispositions are closely related to counterfactuals. One way to understand a disposition such as acceptability is in terms of a corresponding counterfactual: an inference is acceptable by agent Billy just in case, if the inference were made, Billy would accept it. In that case, the Inference Test employs a counterfactual on both sides of its biconditional. We might ask: How can a method guide us concerning the truth of counterfactuals, if it does not tell us how to eliminate them? It appears that we need an antecedent grasp on counterfactual truth to apply the test.

The anxiety is misplaced. The Inference Test is an epistemic method, not a theoretical analysis. So it does not matter if concepts on one side of the biconditional reappear on the other, provided that we have a firm grasp on at least one side at a time. If we tend to have firm intuitions about our actual inferential dispositions, then we can apply these to counterfactual scenarios as easily as to actual ones. There is no difference between assessing an inference

which we *would* make in some counterfactual scenario, and assessing an *actual* inference which one has witnessed or has been told about, or which one expects to occur.[3] We often have firm intuitions about our *actual* inferential dispositions, surely, for we often *actually* make and assess inferences. By applying these actual standards to counterfactual scenarios, the Inference Test can guide our beliefs about counterfactuals, even though it requires a prior intuitive grasp of some counterfactuals.

Does the Inference Test constitute a semantic theory of counterfactuals? We will consider the question in more detail in 3.4.4. But let us be clear that, despite the biconditional, the Inference Test is not supposed to provide truth-conditions for counterfactuals. It does not have the status of a metaphysical or semantic theory. Rather, it has the status of a coherence condition on our beliefs. The Inference Test is normative; as such, agents might as a matter of fact violate it. The claim is that it is incoherent to believe that $A > C$ if you do not accept that one may infer C from A (in the circumstances in which you think $A > C$ holds). Likewise, it is incoherent to accept that one may infer C from A unless you believe that $A > C$ (in the circumstances in which you accept that the inference may be made). This is a test which we can apply to see whether our beliefs in counterfactuals are coherent with our inferential dispositions. It does not provide a hotline to counterfactual truth.

It might be objected that, on its own, A will rarely license an inference to C. This is an important feature of the Test, however. The inferences we will make depend on context. Counterfactuals likewise depend on context. The Inference Test asserts that the acceptability in context of counterfactuals and inferences vary together. Instead of a asserting that biconditional, we could say: $A > C$ in and only in the circumstances where C may be inferred from A.

If the Inference Test is merely a coherence condition on our beliefs, then how can it show anything about counterfactual truth? — By tying our rather uncertain direct intuitions about counterfactuals to a much more robust set of intuitions. Indeed, our inferential dispositions are only intuitions in the philosophical sense, of statements which are to be accepted without argument. In no other sense are they mere intuitions: they are not hunches. We use our ability to form and assess inferences incessantly. Our survival, as individuals and as a species, may frequently depend upon it. For this reason, we may

[3]Williamson deploys a similar line of reasoning in his discussion of expectations, where he talks of our expectation-forming capacities being employed "offline" [Williamson, 2007].

be confident in conferring truth or falsity upon a counterfactual as a result of applying the Inference Test. We could also reason the other way, of course, to accept or reject inferences on the basis of counterfactuals; and perhaps we sometimes do. It is a matter of judgement in each case. But in the cases we shall be considering, the counterfactual intuition will invariably be much weaker than the inferential one. In such cases, the Inference Test will be very helpful.

It remains to justify the claim that the Inference Test represents a genuine coherence condition upon our beliefs. To that end, suppose that you buy a strong coffee, and I remark, "When you've drunk that you'll be perky." I am inferring your future perkiness from my belief that you will drink the coffee.

Since counterfactuals and the deleterious effects of excessive caffeine consumption are both playing on your mind, you pause for thought, then reply: "You think that, if I were to drink this, I would be perky?"

"No," I say, "That I deny."

Now it would be natural for you to ask, "Then why do you think I'll be perky?" And if I offer nothing further, then it would be natural for you to regard me as at least a little strange, perhaps due to having had rather too much coffee myself. We can run a similar story to confirm the other direction of the biconditional. You buy a coffee, I assert the appropriate counterfactual, you tell me you plan to drink the coffee and ask whether I therefore think you will be perky, I deny that I think anything of the sort, and so you think I am strange.

I suggest that the reason you would think me strange is that if I am willing to make the inference, then I should believe the counterfactual. It is this directive of common sense or intuitive rationality which underlies the test I have proposed.

The Inference Test clearly shares a lot of the motivation of the Ramsey Test. Both concern rational belief revision. So what is the difference? Chiefly, that the Inference Test says less, because it has a different purpose. The Ramsey Test offers a minimal (and perhaps normative) hypothesis about the psychology behind evaluation and acceptance of conditionals. The Inference Test merely asserts that we should accept counterfactuals and corresponding inferences under exactly the same circumstances, without specifying what those circumstances are. In this way, the Inference Test avoids specifying principles of belief revision: it invokes our abilities to conduct belief revision appropriately, rather than any hypothesis about how this is or should be done. We can

package all the questions about belief revision under the heading "inference". The fact is, in certain circumstances, we do revise our beliefs in the light of certain facts. Moreover, somehow, we are able to simulate this revision under counterfactual suppositions. The underlying motivation for the Inference Test is that, in some cases, we see this much more clearly when we are asked to think about inferences than we do when we think straightforwardly about counterfactuals.

Perhaps the Inference Test will not always work. There might be counterexamples. It would be better if there were no counterexamples, but if there are, that need not matter. A test for measles need not be perfectly reliable to be useful: if the test is sufficiently cheap and easy to perform, it might still play a part in routine medical tests. High risk groups might be subjected to further tests in the light of negative results, if a false negative is feared. And further tests might follow a positive result, if a false positive is feared. Few medical tests are error-free in either direction, but that does not stop them being useful. A similar usefulness is the goal of the Inference Test. It is not proposed as an analysis, but as a heuristic, a guide, something else, apart from possible worlds, to think about when we are trying to assess counterfactuals. Even if there are counterexamples, the Test might still be useful in this way, so long as we are satisfied that the cases where we use it are not themselves counterexamples to the Test. And as we shall see, the cases with which we shall be concerned are fairly straightforward.

3.4.4 Implications of the Method

An obvious question about the Inference Test concerns the relation between the Inference Test and a semantic theory of counterfactuals. Does the Inference Test imply an inference-based theory of counterfactuals, and does it rule out a world-based semantics? Let us take the questions in turn.

An inference-based theory of counterfactuals, such as Goodman's, seeks to identify counterfactuals with deductive inferences. Well-known problems beset such projects, especially concerning the choice of further premises needed to make deductively valid inferences (cf. Goodman 1983, Kvart 1986, Edgington 1995). Whether or not they can be overcome, it should be clear that the Inference Test does not commit us to such a theory: the normative standards it invokes are not specified, and thus need not be limited to standards of deductive reasoning. On the other hand, a theory which sought to reduce counterfactuals to inductive (that is: non-deductive) inferences would face

the significant challenge of specifying standards of inductive inference. The Inference Test does not commit us to such a reduction, and so does not need to take up those challenges.

Some theorists have been tempted to analyse in the other direction, and appeal to counterfactuals to analyse inductive inference. The Inference Test is entirely compatible with analysing inductive inference with counterfactuals. Inductive inferences might be characterised by supporting certain counterfactuals, for instance, a "tracking" counterfactual stating a counterfactual dependence of evidence upon the truth of a given hypothesis. Lipton proposes the following as a rough characterisation of a good inductive inference:

> ...a strong inductive inference is one where, had the conclusion been false, you would not have made the inference...
>
> (Lipton 2000, 185)

This is intended as a simple characterisation for further refinement, but the refinements will not make any difference to question of whether counterfactual accounts of inference are compatible with an inference-based method for guiding our beliefs about counterfactuals.

On the one hand, Lipton's proposal makes a good inductive inference a sufficient condition for the truth of the corresponding counterfactual. It is, therefore, well aligned with the claim that accepting an inference imposes a coherence requirement to believe the corresponding counterfactual. The Inference Test says that if you accept the inference, you should accept the counterfactual, and Lipton's tracking condition makes the counterfactual true in circumstances where the inference is a good one. And if you accept the inference, you presumably think it is good. So the tracking condition seems to fit with our directive, since the directive says you should believe what the tracking condition asserts.

On the other hand, the Inference Test further asserts that the inferibility (in the sense recently outlined) of consequent from antecedent is a necessary condition on our coherent acceptance of a counterfactual. If we accept the counterfactual, we must accept the inference. In order to be incompatible with the Inference Test, Lipton's tracking condition would further need to deny that inferibility of consequent from antecedent is a necessary condition on our coherent acceptance of a counterfactual. The tracking condition on its own clearly does not constitute or imply any such denial. The compatibility appears likely to generalise to other accounts, but even if there are exceptions,

compatibility with just one analysis of inductive inference in terms of counterfactuals suffices to show that the Inference Test is not in principle incompatible with such analyses.

A related but opposite worry is that the Inference Test might be incompatible with a counterfactual account of inductive inference, not by being contradictory, but because it might make such an account circular. I hope that the status of the Inference Test as a test, and not an analysis, is sufficiently clear to avert this worry. The following picture is not circular: good evidence is characterised by counterfactual reliability; and to find out the truth of a counterfactual, we can ask whether we would accept a corresponding inference. An analysis of inference in terms of a something more basic, counterfactual dependence, is quite compatible with the thought that we might find out about the more basic thing via the inferences it is used to analyse. Indeed, we might even expect some such situation, just as our primary acquaintance with some of the cruder, collective properties of atoms is through the medium sized dry goods which they are thought to underlie.

So the Inference Test neither entails an inference-based analysis of counterfactuals, nor rules out a counterfactual analysis of inference. However, it is not entirely neutral with respect to semantic theories for counterfactuals. When those theories make claims about the truth of counterfactuals, the Inference Test will either confirm or deny the theory making those claims, depending on whether it agrees with the claims. There are two ways this could prove problematic.

First, a semantics might ground a logic of counterfactuals which the Inference Test does not fully reflect, either because it disagrees directly with certain moves or because the Test is just too coarse to apply meaningfully to complex moves. I do not think this is a serious problem. The Inference Test can admit of exceptions, as we have seen. A semantic theory might also admit of exceptions if it is understood, as such a theory must surely be, as an *idealisation* with respect to common practices. Among those practices might be the application the Inference Test.

Second, a semantics might disagree with some substantive claim about the non-logical properties of counterfactuals. It is evident that we will encounter problems of this sort:

> ...those who are adverse to backtracking counterfactuals may have to analyse reliability with different conditionals, because of the prevalence of inductive inferences that move from past to future.

(Lipton 2000, 185)

Obviously the Inference Test will be incompatible with a ban on backtrackers for a similar reason. We make inferences in both temporal directions. So counterfactuals must be true in both temporal directions.

At this stage, I hope this will be seen as an advantage. The Inference Test lacks temporal bias. It is in an excellent position, therefore, to tell us something about the truth of the matter concerning the asymmetry of counterfactual dependence. In the next chapter, we shall discuss causal selection. I shall argue that the Inference Test reveals an asymmetry of counterfactual dependence which is much weaker than Lewis's. Some backtrackers are true, but they are fewer than true foretrackers. In the next chapter, I shall use the Inference Test to argue that the circumstances in which backtrackers are true correspond to the circumstances in which we select *the* cause from the mere conditions for an effect.

3.5 Summary

We began by considering the plausibility of the idea that backtrackers are false under normal resolutions of the vagueness infecting all counterfactuals. Two arguments for this claim were criticised. The first was that mixing backtrackers and foretrackers leads to contradictions, because they are generally true under different resolutions of vagueness. It was accepted that care must be taken when mixing backtrackers and foretrackers, in order to avoid flipping between different resolutions of vagueness; but it was argued that this point applies equally to mixing foretrackers alone, and does not suffice to show that backtracking and foretracking contexts are distinct. The second argument concerned the grammatical complexity associated with backtrackers, and the corresponding strangeness of backtrackers which are forced into the grammatical subjunctive which is so natural for expressing foretrackers. I suggested that the falsity of backtrackers could not explain their odd sound, and that the complexity of backtrackers did not imply a shift of context to favour a special backtracking resolution of vagueness. Thus both reasons for thinking that backtrackers and foretrackers were true under different resolutions of vagueness were rejected.

This line of argument is intended to deprive Lewis's asymmetry of counterfactual dependence of its explanandum. If there is no reason to suppose that there is an ordinary, foretracking context, then there is no reason to seek an

account of such a context — and thus no reason to argue for a sharp distinction between true foretrackers and false backtrackers. This line of argument is also directed at the more general worry, that even if Lewis's asymmetry thesis fails for internal reasons, some asymmetry of counterfactual dependence nevertheless exists. I did not argue that there is no asymmetry; rather, I argued that there is no reason to think that backtrackers and foretrackers must be kept apart, each to their own special context.

It might be objected that Lewis's asymmetry of counterfactual dependence was fundamentally motivated by explaining more general puzzles, which we have not considered at all: in particular, the perceived "openness" of the future contrasted with the fixity of the past. This is, in part, because the contrast is not easy to get a grip on; nor is it obvious what an account of it would be like. Lewis's asymmetry of counterfactual dependence fails to provide a very convincing account, because — as we saw in 2.4.1 — Lewis's account allows some backtracking. Yet it is not as if we regard the past as mostly but not entirely fixed, or variably fixed depending on context. We may *hold* it fixed variably with context, when we are entertaining various counterfactual suppositions; but that is quite different from *regarding* it as fixed in a variable and not-quite-complete way.[4] This qualm aside, it is far from clear that a very general puzzle such as the asymmetry of openness directly motivates any very specific explanation, such as Lewis's asymmetry of counterfactual dependence; the asymmetry of openness might be compatible with other explanations. The other important asymmetries which Lewis had hoped to explain are the temporal asymmetry of causation, and the direction of time itself. These questions are largely beyond the present scope, but will be touched upon in Chapter 7.[5]

Having argued that the existence and nature of an asymmetry of counterfactual dependence is an open question, not decided by considerations of logic or of English grammar, we considered methods for assessing counterfactuals without temporal bias. Lewis's method was criticised — perhaps a little unfairly, since it is not advanced as a method. But the method is in widespread (if implicit) use, I fear: if so, identifying it and its flaws is important; and if not, the only injury done is to a straw man. The Ramsey Test was found to be insufficient as a test for counterfactuals, because its results

[4] In this point I am indebted to Peter Lipton.

[5] It might further be objected that we need an asymmetry of counterfactual dependence for other purposes, for example, in causal decision theory (cf. Gibbard and Harper 1978, Lewis 1981, Elga 2000). Such a need ought, however, to be subservient to the prior question of whether there is in fact an asymmetry of counterfactual dependence. If not, then applications which depend on there being an asymmetry will need to be revisited.

are not determinate until a heuristic for belief revision is inserted. However Ramsey's suggestion that the move from antecedent to consequent be seen as an inference was picked up, with help from Edgington and Goodman, and the Inference Test was proposed. It was defended as a coherence condition on our beliefs: accept the counterfactual, and you must accept the inference from antecedent to consequent. A cursory defence against obvious objections was conducted, and the Test was presented as compatible with world-based semantics for counterfactuals and with counterfactual analyses of inference.

Now we turn to the main topic: causation. However the discussion of causation, particularly in the next chapter, can also be seen as an exploration of the real asymmetry of which Lewis's thesis of the asymmetry counterfactual dependence is a parody. We will be exploring the circumstances under which backtrackers are true. In Chapter 4, I shall argue that the circumstances in which a backtracker is true are just those in which we distinguish the cause from the mere conditions for a given effect.

Chapter 4

Selection

4.0 Abstract

I strike a match, and it lights. It is unusual to say that the presence of oxygen caused the flame, even though we might be fully aware that oxygen is needed for the flame. This chapter presses the problem of selecting the cause from among mere conditions, which has often been dismissed by theorists of causation as a sort of whim. I start by arguing that selection needs to be accounted for, and sketching some of the approaches which have been tried. I focus on the strategy of assimilating causal selection to the contrastive mechanism of causal explanation. Although this approach enjoys some success, I argue that it fails either properly to explain or to enable us to justify our selective practices. I introduce the *Reverse Counterfactual* which, I suggest, captures the intuitive notion that causes *make the difference* to their effects. I argue that the Reverse Counterfactual is true of causes but not of mere conditions. This yields an account of selection which overcomes the objections raised against the contrastive strategy, and which links the context-sensitivity of causal selection to the context-sensitivity of counterfactuals. The relation between contrastive explanation and the Reverse Counterfactual is discussed, along with various objections.

4.1 The Problem of Causal Selection

Here are six propositions.

(1) Sometimes we say of the cause, but not of any mere condition, that it is the cause of a given effect.

(2) Sometimes we cite the cause, but not any mere condition, in explanation of a given effect.

(3) Sometimes we use the cause, but not any mere condition, to predict a given effect.

(4) Sometimes we seek to bring about the cause, but not any mere condition, in order to bring about a given effect.

(5) Sometimes we make moral judgements on the basis that an agent's action is the cause of a given effect, which we would not on the basis that it was a mere condition.

(6) Sometimes we make legal judgements on the basis that an agent's action is the cause of a given effect, which we would not on the basis that it was a mere condition.

If any of these are true, then a question arises: what is the difference between a cause and a mere condition? The *problem of causal selection* is the fact that this question is hard to answer adequately. However, many philosophers of causation do not think that selection poses a particular problem; they do not acknowledge the problem of selection as such. For a notable example, David Lewis developed his theory of causation extensively in an effort to deal with the problem of preemption, but no changes were motivated by giving an account of selection.

The reason is that Lewis did not believe that an account of the selection of cause from mere conditions belonged in an account of causation. For he did not believe that causal selection is strictly a feature of causation itself.

> We sometimes single out one among all the causes of some event and call it "the" cause. 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... I have nothing to say about these principles of invidious discrimination. I am concerned with the prior question of what it is to be one of the causes (unselectively speaking). My analysis is meant to capture a broad and non-discriminatory notion of causation.
>
> (Lewis 1986a, 162)

Lewis's analysis does indeed capture a non-discriminatory, unselective notion, for on all his counterfactual accounts, there is no distinction between cause

and mere condition. The starting thought is that c causes e if $(\sim C > \sim E)$. Sitting here at a desk in a warm dry room, I strike a match, and it lights. If the lighting of the match is e, then clearly, on this embryonic Lewisian analysis, the match strike and the presence of oxygen have an equal claim to be c. For without either, the match would not have lit. This feature is preserved through all Lewis's considerable elaborations of the starting thought. On all of Lewis's accounts, the extension of the term "mere condition" would be just the same as the extension of "cause", disregarding conversational implication and any other conventions. Even if we sometimes discriminate between causes and conditions, and mark the discrimination by calling some event "*the* cause", we discriminate on some other basis than the instantiation of some general causal difference between cause and condition. On Lewis's view, we distinguish cause from condition for reasons that are entirely our own.

The Lewisian view of causal selection is shared by another famous account of causation which is otherwise very different. Mill was similarly hostile to selection as a feature of causes themselves, despite his interest in experimental methods for singling out causal factors. The reason was that Mill considered causation a *law*, by which he meant an exceptionless regularity. Indeed Mill thought that the law of causation was the only exceptionless regularity.[1] This led him to the doctrine of the "whole cause", by the following route. We might say that Jones's eating a dish caused her death, on the basis that if she had not eaten it she would not have died. But there is not an exceptionless regularity between the eating of the dish and death, unless we specify the eating of the dish so precisely that only Jones's eating counts. Others with a stronger constitution, or lucky enough to receive prompter and better medical care, eat the same dish or another just like it, and survive. So strictly speaking, Jones's health, the other parts of her meal, and "perhaps even the present state of the atmosphere" (Mill 1887, 237) jointly count as the cause, because only when taken together are these exceptionlessly followed by the effect.

> The real Cause, is the whole of these antecedents; and we have, philosophically speaking, no right to give the name of cause to one of them, exclusively of the others.
>
> (Mill 1887, 237)

[1] "Now among all those uniformities in the succession of phenomena, which common observation is sufficient to bring to light, there are very few which have any, even apparent, pretension to this rigorous indefeasibility: and of these few, one only has been found capable of completely sustaining it" (Mill 1887, 235). That one is the law of causation.

In short, the dominant view across a broad spectrum of Humean accounts denies that there is any difference between "the" cause and the mere conditions for an effect (apart from our differential treatment) — either because each condition is a cause (Lewis) or because the cause is all the conditions taken together (Mill). But if any of (1)-(6) is ever true, then it follows that there is *some* difference between cause and mere condition, in particular cases where some of (1)-(6) are true. So on non-selective accounts of causation like Mill's and Lewis's, any general difference between causes and conditions cannot be objective; and the many objective differences which presumably exist between a given cause and condition must be particular to the case in hand, and cannot be general.

If (1)-(6) are not accounted for by a theory of causation, then they must be accounted for (if they are to be explained at all) by appeal to something other than a theory of causation. Yet the account ought also to throw some light on the fact that we treat unselective causation with these selective foibles. In fact there are two needs for such an account. The first is the obvious one: an account of the principles governing causal selection is needed if we are to understand those principles. The second need has been given less attention, but deserves more: namely, the account should explain why we insist on *using* causation as a selector in the way that we do, when causation is allegedly so ill-suited to the task. Why do we appeal to causal concepts in order to distinguish the match strike from the presence of oxygen, the arsonist's action from the builder's, and so on, across the contexts of (1)-(6), if there is no general difference between the event we distinguish and those from which we distinguish it in those cases? On an unselective account of causation, we systematically misuse our causal concept by using it to select causes from conditions: we make what Lewis calls an "invidious discrimination", which according to Mill we have "philosophically speaking, no right" to make. Such a view amounts to a sort of error theory about much of our ordinary causal talk; error theories can be correct, but they generate an explanatory need — the need to explain why we have fallen into systematic and widespread error.

This is a *prima facie* objection to the prevailing view, that an account of causal selection does not belong in an account of causation. I am aware of no serious efforts answer this objection, and thus to meet this second explanatory need we have identified. No proponent of an unselective view of causation seems to have devoted much thought to explaining why we use it selectively. Efforts have been made, however, to give accounts of the principles governing

causal selection. I shall seek to present existing work on causal selection as a sort of progression through three increasingly sophisticated strategies. Then I shall consider the most sophisticated of these in detail, and argue that it answers neither of the needs just identified. Finally I shall propose and defend an account which I hope is more successful, and try to explain the limited successes of previous accounts.[2]

4.2 Three Steps Towards an Account of Selection

4.2.1 The Special Event Strategy

Perhaps the most obvious way to try to account for causal selection is to seek to identify the kinds of events which we tend to select as causes. This approach has the advantage of pragmatism, since no great theory is required to observe which kinds of events we identify in which circumstances. Hart and Honore adopt this strategy. They suggest that we distinguish two sorts of events from the general background: *human interventions* and *abnormal events*. It may be hard to give a precise account of these, but as lawyers their interests are pragmatic, and although it may be hard to account for, the fact is that, in practice, we can usually identify both human interventions and events we consider to be abnormal relative to the usual course of things.

Hart and Honore note that selection is *flexible* (my term), in the sense that we select different events as the cause in different circumstances. They differentiate two ways in which selection depends on context. Consider an explosion, requiring fuel, oxygen and ignition. If it occurred in a petrol station, we would usually say that it was caused by the ignition — the dropping of the cigarette, the spark, or whatever it was. We would not usually say that it was caused by the presence of oxygen. This choice, however, is flexible in two ways. First, of that very explosion, we might sometimes admit that the oxygen is a cause, if the context changes suitably: for example, if we are discussing the chemistry of explosions, or the possible measures for preventing them. Thus selection of the cause is relative to the *context of inquiry* — the context in which we discuss, ask about, investigate or mention the effect. Second, if a similar explosion occurred not in a petrol station but in a factory, following a

[2]One area where causal selection is currently topical, which I shall not discuss but which could benefit from philosophers taking causal selection more seriously, is epidemiology (cf. Whitbeck 1977, Krieger 1994, Parascandola and Weed 2001), .

leak in a process normally conducted in a vacuum, we would probably identify the leak of air into the process as the cause. Thus selection of the cause is relative to the *context of occurrence* — the time and place where the effect occurred, the frequency of such events, and properties of the wider situation more generally. (Adapted from Hart and Honore 1985, 10.)

Hart and Honore's discussion usefully emphasises the complexity of our selective practices, identifying and clarifying flexibility as an important feature which any account of selection must accommodate. Their book also serves to emphasise the weight we put upon causal selection — the seriousness with which we regard the difference between cause and condition. After all, their discussion concerns the place of causal selection in the law, and a lot that we value can hang on legal decisions. This is a point to which we will return in 4.3.2 below.

However their approach suffers from a number of drawbacks if it is seen as a philosophical theory of causation. It does not have the metaphysical credentials we would expect of a modern theory of causation. Its form is disjunctive, its two central concepts (human intervention and abnormal event) are both vague, and the resulting theory would be highly interest-relative. Hart and Honore advance their theory for legal, not metaphysical, purposes, so perhaps this is not surprising. Perhaps, then, we should see theirs as an account of our selective practices, independent of any account of causation.

But then the account clearly will not answer the second of the questions we set out at the beginning of the chapter. It will not explain why we use causation — causal concepts, causal language, causal reasoning — to select causes from mere conditions.

Moreover, it is far from clear that it will answer the first, descriptive, question adequately from a philosophical point of view. Describing our selective practices is not the same as offering an account of them for philosophical purposes. Mere description may be useful for clarifying legal argument, but it will not answer the obvious philosophical question: it will not explain *why* we single out human acts and abnormal events. And until the principles underlying our preference for these sorts of special events have been given, there is a high risk of counterexamples (cf. Beebee 2004).

For these reasons, then, the special event strategy is at best a starting point for a principled account of selection. It may offer useful rules of thumb for legal contexts, but it seems neither sufficiently explanatory nor sufficiently reductive to amount to an enlightening philosophical theory.

4.2.2 The Causal Field Strategy

Mackie also toys with the special event strategy, and comes to similar conclusions to Hart and Honore about the sort of events we tend to select as causes. However he proposes an explanation for our tendency to select those sorts of events. Mackie's account starts with the thought that all causal statements are responses to *causal questions*, and that causal questions are to be understood as questions about what *makes the difference*:

> A causal statement will be the answer to a causal question, and the question 'What caused this explosion?' can be expanded into 'What made the difference between those times, or those cases, within a certain range, in which no such explosion occurred, and this case in which an explosion did occur?'
>
> (Mackie 1974, 35)

In answering such questions, Mackie thinks we rule out irrelevant events in two ways. First, such questions are asked and therefore answered against the background of a *causal field*, which is simply a set (he says *range*) of possible cases sharing some features with the actual case. We ignore all the many possible ways in which the effect could have been prevented, except those which are differences between the actual case and another case in the causal field. For example, we ignore the possibility that the Earth never came into being, that the factory was never built, that the town was flooded in a seismic upheaval, and so on when we consider what caused the explosion: we are looking for differences between the case where the explosion occurred and cases where it did not — but only *within a restricted field* of cases. These cases are alike in respect of the existence of the Earth, the presence of a factory, and the absence of encroaching seas. So we do not identify these factors as causes, even though we agree that they are counterfactually necessary for the effect.

Relativising causal questions to a restricted field does not, according to Mackie, always narrow things down to just one difference:

> Any part of the chosen field is decisively ruled out as cause; a more elusive point is that among factors not so ruled out which are severally necessary for the effect, we still show some degree of preference...
>
> (Mackie 1974, 35)

The more elusive part of our selective judgements reflects our interests in particular circumstances, Mackie thinks. For example:

> ...we may agree that the collision was caused just as much by Smith's driving straight ahead as by Brown's deviating to his right without warning, but say that it is more important, for moral and legal purposes, to draw attention to the second of the two causal relationships.
>
> (Mackie 1974, 36)

His position is that causes are those differences relative to a particular causal field, and that interests further select *among* these causes, when mention of the fact that a given event is a cause "happens to be irrelevant" (1974, 36).

Subsequent work has not paid much attention to Mackie's strong distinction between these two elements in causal selection — namely, relativity to a causal field, and interest-relativity. But Mackie is explicit that he considers them both to be important. He thinks that we could do away with either component of the account altogether — making selection either purely interest-relative or purely relative to a causal field — but that in fact both elements are present in our thinking (cf. Mackie 1974, 36).

The causal field is not absolute for any given cause-effect pair; rather, it varies with the context. In particular, it varies depending on what causal question (perhaps implicit) a given causal claim answers (Mackie 1974, 35). An obvious objection would deny that all causal claims are responses to implicit questions. We will discuss this objection when we consider the contrastive strategy (4.3.1). On the other hand, the suggestion that causal statements are answers to implicit questions enables Mackie to account for the flexibility of selection, relative to the context of inquiry. Obviously, the answer we give depends upon the question we ask: so when we ask, "Why did this explosion occur?", we might implicitly be asking, "What made the difference between this explosion and other, non-exploding petrol stations?" The other petrol stations we have in mind are well-supplied with oxygen, so we pass over the oxygen as a possible cause of the explosion. However, the answer will be different when the implicit question is, "What made the difference between this explosion and other occasions when this delicate manufacturing process did not blow up?" Oxygen may have been absent on those occasions, so we would not rule it out as a cause (though as previously explained, whether we mention it and thus positively rule it *in* depends further on our particular interests in the case).

So causal fields explain why selection is relative to the context of inquiry, if they are determined by implicit questions as Mackie suggests. However it is less clear that the causal field approach explains how selection depends on the context of occurrence, at least as Mackie develops it. Causal fields are defined by the implicit questions to which causal statements are allegedly an answer. This makes it clear how selection can depend on inquiry. But it is not clear how selection depends on the context in which things happen, since it is not clear why certain events prompt certain sorts of questions. It may be perfectly obvious that an explosion in a petrol station suggests a different question from an explosion in a factory. But why? Even assuming that it is clear how a particular question generates a particular causal field, it is not clear how a particular context generates that question. The relation between context of occurrence and context of inquiry is not explained. Yet Mackie's account of selection makes it entirely dependent on the context of inquiry; so if his account is to explain how selection depends on the context of occurrence, it will need to explain how our inquiries depend on their subject matter. That is a difficult and strange task; usually we suppose that we can investigate what we want, and that choices about what question to ask are entirely pragmatic, or interest-relative. It seems, then, that in the end, Mackie's account must make the dependence of causal selection on the context of occurrence pragmatic: as pragmatic as our choice of what question to ask. This is not necessarily a criticism, but it serves to show that — despite the machinery of causal fields — Mackie's position is still fairly close to that of Mill and Lewis.[3]

4.2.3 The Contrastive Strategy

Another unclarity in Mackie's suggestion concerns the way in which causal fields are generated by questions: what is the precise nature of those questions, and how do they determine a causal field? It seems to me that contrastive accounts of selection can be seen as refining this aspect of Mackie's account (sometimes explicitly, as in the case of Schaffer [2005]). Mackie's own examples, such as that of the explosion (given previously on page 73), are expansions of causal questions into *contrastive why-questions*: questions of the form, "Why X rather than Y?" This suggests that contrasts play a central role in determining what Mackie referred to as the causal field. The contrastive

[3]A concise formulation of Mackie's account of causal selection can also be found in [Mackie, 1965]. Two recent and quite different attempts to develop a Mackian position are [Menzies, 2004] and [Schaffer, 2005].

accounts we shall consider drop the "causal field" terminology, but their effect is the same: contrasts determine a set of cases sharing certain features, thereby narrowing our attention to the handful of differences remaining. Contrastive accounts may therefore be seen as deepening Mackie's approach, by replacing the blunt notion of a causal field with the notion of a contrast. Contrasts, unlike causal fields, are things we are already familiar with. Moreover the contrastive account of selection is more specific about the mechanism by which our causal claims pick out a set of possible cases, against which we compare the actual case for differences. In both respects, it constitutes a superior explanation of causal selection, while yielding a theory of roughly the same shape.

We shall focus on the contrastive strategy. It is probably the best-known style of account of causal selection. This may be partly because it can be implemented without commitment to any particular theory of causation. Thus it typifies the style of account which the Mill-Lewis line of thought recommends. It is also a promising kind of account, enjoying some remarkable successes, as we shall now see.

4.3 The Contrastive Strategy

4.3.1 Contrastive Explanation and Causal Selection

Lipton hopes to shed light on causal selection by appeal to contrastive causal explanation, and he does so without endorsing any particular theory of causation. Moreover, he has also given explicit attention to other items (apart from explanation) on the list (1)-(6), notably selection in legal contexts. This is a context where explanation is often not the explicit goal, so it offers a good test for the strategy of assimilating causal selection to the contrastive mechanism claimed for causal explanation.

The central piece of Lipton's account of contrastive explanation is his *Difference Condition*, offered as a necessary condition on explanation:

> *To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the case of not-Q.*
>
> (Lipton 2004, 42 — emphasis original)

For example, suppose Able and Baker both propose to Suzy, who accepts Able. Suzy's mother asks her why she accepted Able (P) rather than accepting Baker

(Q). Suppose that Suzy says, "Because Able proposed." That might be a good explanation, if Baker did not propose. But since Baker also proposed, then clearly the fact that Able proposed is no explanation at all of Suzy's preference for Baker. We have an event — Able's proposal — which is indeed a cause of the acceptance (P), but which fails Lipton's Difference Condition in virtue of the presence of a corresponding event — Baker's proposal — in the case of Baker's rejection ($\sim Q$). On the other hand, if Suzy said, "I'm marrying Able because he is a devastating logician," then we might have an explanation: for Baker is not a devastating logician. Thus we have a cause of Able's acceptance (P), namely Able's logical prowess, which is absent from Baker's rejection ($\sim Q$), because Baker is logically mediocre. The Difference Condition is, however, only a necessary condition on explanation, and I hope the example illustrates this too: Suzy's mother is unlikely to accept Able's logical prowess in explanation of Suzy's choice, even though it meets the Difference Condition. Nevertheless, logical prowess cannot be excluded as an explanation of Suzy's choice in the decisive way that we earlier excluded Able's proposal, which failed the Difference Condition.

Lipton's proposed strategy for dealing with causal selection is to claim that selection of the cause employs the contrastive mechanism claimed for this sort of explanation:

> ...a cause marks a difference between the situation where the effect occurs and a contrasting situation where it does not.
>
> (Lipton 1992, 136)

This yields a contrastive account of selection. For each of (1)-(6), we say of the cause, and not of any mere condition, that it is the cause, because the cause is a difference between this case where the effect occurs and contrast cases where it does not.

One obvious problem is that, in general, there may be many differences between the history of a given effect and the history of a given non-occurring contrast. This is a problem inherited from the account of contrastive explanation from which this account of causal selection is descended. Lipton therefore acknowledges the need for further principles.

> ...the contrasts we construct will almost always leave multiple differences that meet the Difference Condition. At the same time, however, some causally relevant differences will not be explanatory in a particular context, so while the Difference Condition may

> be necessary for the causal contrastive explanations of particular events, it is not generally sufficient. For that we need further principles of causal selection.
>
> (Lipton 2004, 47)

Lipton duly provides a discussion of these further principles. Contrast is not the whole story. Nevertheless, just as the Difference Condition is offered as a necessary condition on explanation, so it might be for causal selection. We might consider the contrastive strategy as indicating an important necessary condition for an event to qualify as the cause, even if the condition is merely necessary, not the whole story. For this reason I am not inclined to regard this problem as a very serious one: accounting for selection in terms of contrast might still be illuminating even if the account is only partial.

Since Lipton's account of selection is a descriptive analysis of common practice, the argument for it proceeds largely by exhibiting examples where the cause indeed appears to be the difference between the case where the effect occurs and some contrast; and, of course, contesting any counterexamples. I do not propose that we debate the success of the contrastive account in this respect: the idea is plausible and thoroughly discussed elsewhere (cf. Lipton 2004, Schaffer 2005), and my interest is in the prospects for this strategy of accounting for selection *if* contrast can be shown to line up appropriately with our selective intuitions in a significant range of cases. I shall argue in 4.3.2 below that even if the cause marks a difference between the effect and some contrast case, a contrastive account of selection encounters serious difficulties.

One misplaced strategic objection to the contrastive approach should be set aside. Menzies objects that "one of the central assumptions of the strategy — that every causal statement must be understood in the context of an implicit contrastive why-question — is too strong" (Menzies 2004, 150). This is an understandable reaction, especially given that, as we saw, Mackie's account of selection explicitly endorses the view that every causal statement is an answer to an implicit why-question. But I think Menzies' objection is misplaced, because the contrastive strategy need not endorse this view. Note that we have described the contrastive strategy without hinting at implicit why-questions. The strategy requires that there be a contrast for the effect when we select the cause from among the mere conditions, but it does not require that there is an implicit *question* about why the effect rather than the contrast occurred. The claim is not that causal selection *is* contrastive explanation, but merely that it shares the contrastive mechanism. The assumption that selection is relative to

some contrast for the effect in question is a weaker assumption, than that causal selection always implies a why-question. It is also more plausible, because of the familiar flexibility of selection: in different contexts, we cite different events as the cause of the same effect. I might say the lasagne was good because I used more tomatoes, but my dinner guests would say it was because I am finally learning to cook. This flexibility would be explained neatly by the contrastive strategy, because different contexts might yield different contrast choices, and thus different selections. Although why-questions are a source of contrast, there is no reason to think they are the only source, and thus no reason to tie the flexibility of selection to the vigour of our enquiries.[4]

4.3.2 Two Objections

Here come two objections to the contrastive account of causal selection, both arising from difficulties about how the choice of contrast is determined. Lipton argues that any such objection to his account of explanation is ill-founded:

> My goal... is to show how the choice of contrast helps to determine an explanatory cause, not to show why we choose one contrast rather than another. The latter question is not part of providing a model of explanation, as that task has traditionally been construed. It is no criticism... of my account of contrastive explanation that it does not tell us why we are interested in explaining some contrasts rather than others.
>
> (Lipton 2004, 46)

This response may be appropriate for an account of explanation, but not, I shall argue, for causal selection, for two reasons. First, *explaining* causal selection requires us to explain, not just why a certain cause is selected given a certain choice of contrast, but why that contrast was chosen in the first place. Second, *justifying* moral and legal decisions which involve causal selection requires, not just saying why a certain cause is appropriate given a certain contrast, but justifying that choice of contrast in the first place.

[4] I think a parallel response could be made in defence of Mackie's position. Although Mackie commits himself to the claim that all causal statements are responses to implicit why-questions, the notion of causal field does not require this. Menzies develops a causal field account of selection which does not make causal fields relative to implicit why-questions, as we shall see in 4.6.3. Menzies has three other objections to the contrastive strategy [Menzies, 2004]. Two are specific to Lewis's account of contrastive explanation, and do not apply to Lipton's. The remaining one we shall consider shortly.

The explanatory shortfall of the contrastive account of selection may be brought out in various ways. Menzies puts it like this:

> ...the two-part strategy [of giving an account of causation and then tacking on a contrastive account of selection] is unsatisfactory from an explanatory point of view. It unnecessarily duplicates the idea of a cause as something that makes a difference: first in the analysis of "objective cause" as something that makes a counterfactual difference; and then again in the contrastive explanation account of the "context-sensitive cause"... [S]urely it would be a surprising fact, requiring elaborate explanation, if our framework for conceptualizing causation used in two different but crucial ways the very same idea of difference making. It would be much more likely that our conceptual framework was developed on the basis of a single fundamental application of this idea.
>
> (Menzies 2004, 150-1)

Menzies is relying on a platitudinous notion of cause as something which "makes a difference" to its effect. He takes it that this notion is central both to the analysis of "objective" causation, and to the analysis of causal selection. If so, he says, then we should seek a unified analysis of causation and causal selection which applies the intuitive idea of difference-making. An analysis which is not unified in this way generates an explanatory need, one that has not been answered.

This objection is a more specific variant of the general *prima facie* objection I put in 4.1, where I suggested that an unselective account of causation generates a need to explain our selective application of causal concepts (besides failing to address the preexisting need for a descriptive account of the principles governing selection). Another specific variant of this line of objection, which is also present in Menzies [2004], is that counterfactuals themselves are context-dependent. Thus context already plays a role in a counterfactual theory of causation. On an unselective counterfactual theory, context comes into play in two quite different ways: first, in fixing the truth-values of the counterfactuals which determine what causes what, and second, in choosing between those counterfactuals when cause is distinguished from mere condition.

Perhaps the most striking way to bring out the explanatory shortcomings of the contrastive approach arises when we remember that explanation and prediction are often symmetric. An explanation of the observed position of

the moon in terms of its roughly Newtonian progress from its position this time yesterday will also be a prediction of the current position of the moon, if it is offered at an earlier time, or at a present or later time in the absence of an observation. This symmetry is not always present, but there is usually an identifiable reason for its absence. In the present case, if the contrastive account is a good explanation of causal selection, we might expect it to make predictions about what we will select in certain circumsances. *But a contrastive account yields no concrete predictions without an account of contrast choice*. For example, it does not tell us that I will select the match strike rather than the oxygen to explain or predict or produce the flame, no matter how much information we make available. For it requires a contrast for the flame before it can get started. This is not a case where the explanation/prediction symmetry breaks down; the contrastive account is not intended as an account of contrast choice. Turning our attention to prediction merely emphasises that, if causal selection indeed employs a contrastive mechanism, then contrast choice must also be explained. For, if indeed it employs a contrastive mechanism, then causal selection *includes* a decision about which contrast is appropriate for a given context. The contrastive account cannot predict what we will select in a given context until we have fed it the contrast appropriate to that context. This, I suggest, is a clear indication that the contrastive account does not fully explain selection.

The second class of objections concerns moral and legal justification for decisions which depend upon the distinction between cause and condition. Causal selection is central, not only in explanatory contexts, but also in moral and legal contexts. The contrastive account derives from an account of explanation, but moral and legal contexts differ from explanatory contexts in a relevant way. We are free to explain what we wish: the context of explanation does not determine what we should explain, only whether an explanation is good given an explanandum. It is far from clear that contexts of moral and legal evaluation are similar in this respect. If an arsonist burns a warehouse to the ground, she is morally and legally condemned, not merely because we happen to take an interest in evaluating her actions, but simply because she has done something wrong and illegal. In this case, as in many others, her crime is a cause: she caused the destruction of a warehouse (as well as satisfying whatever other conditions are necessary for the crime of arson). We may allow that the choice of effects whose causes are to be morally or legally evaluated may be left out of an account of causal selection. But even if we specify independently

that the fire is to be evaluated, the contrastive account still gives too much latitude. If we contrast the burning of the warehouse with the absence of a warehouse at all (eg. in the course of a comparative study of crime patterns in industrial versus residential areas), the action of the builder qualifies as a relevant difference. Yet we imprison the arsonist, and may employ the builder for the rebuild: this shows, as clearly as anything can, that we do not treat the builder as causing the fire. The moral and legal weight that our discrimination between builder and arsonist must bear is substantial, and until we have an account of the way we choose our contrast, the contrastive account does not hold up. It only provides an account of how selection proceeds once we have a contrast. In moral and legal contexts, the choice of contrast matters just as much as the contrastive mechanism, since both determine the facts upon which moral or legal judgement is to be passed. And it is clearly not morally or legally acceptable, as it often is explanatorily acceptable, to leave one of these determinants entirely up to us: it is a matter of fact, not up to us, whether the arsonist or the builder committed a moral or legal crime.

Both my objections may be seen as deepening an objection I sketched against Mackie in 4.2.2. There, I criticised his causal field account for making selection depend entirely on the context of inquiry, and not at all on the context of occurrence. That was because the causal field was generated entirely by an implicit why-question. I have argued that contrasts need not be generated by implicit why-questions. Nevertheless, the contrastive account as it stands does not relate the source of the contrast at all to the context of occurrence. Perhaps you might think that it should not: if we believe, with Lewis and Mill, that selection is something we contribute after the causal facts are fixed, then perhaps an account which makes selection suitably independent of the context of occurrence would be desirable. But even this radical view about the place of selection does not excuse us from accounting for it. Our actual selective practices do appear to depend on the context of occurrence as well as the context of inquiry, as Hart and Honore point out. If we accept that (and surely we should), then even an account of selection which makes it entirely independent of the facts of causation, and entirely dependent on our attitude towards those facts, must explain how the context of occurrence affects our choices. The intention of the two objections I have put is to push this point.

4.3.3 Pragmatics

It is often, I suspect, assumed that both kinds of objection are to be handled by an appeal to pragmatics (cf. Lewis 2004a, Menzies 2004, Schaffer 2005). The explanatory need generated by an unselective account of causation is to be answered, in each of its various forms, by giving an account of the pragmatics of causal selection, and especially of contrast choice. Pragmatic factors might explain why we use causal terminology to make distinctions which are not purely causal, just as pragmatic factors might be thought to explain why we commonly use "if" to imply "only if". In reply to Menzies, perhaps difference-making is a central causal notion, but when we select cause from conditions we are taking into account further pragmatic considerations which are absent when we apply the concept to determine the "objective" cause. Likewise, counterfactuals depend on context to resolve their vagueness and thus arrive at a determinate truth-value; there is nothing wrong with supposing that, even after vagueness has been resolved, pragmatic considerations might make some true counterfactuals more salient, more relevant or more appropriate to mention in some contexts than others. I suggested that the contrastive account's lack of predictive power illustrated lack of explanatory power, but pragmatic accounts are a standard class of exceptions to the prediction/explanation symmetry. Even if general principles governing the pragmatics of a certain practice are given, the principles rarely suffice for predictions, because pragmatic considerations are usually complex, subtle and generally resistant to fully reductive explanation.

The objection from moral and legal justification might likewise be thought to succumb to a pragmatic treatment. Clearly, in moral and legal contexts, we have interests which might be absent in the case of explanation. Perhaps, then, the contrastive account lifted from the explanatory context is indeed more basic. The interests we have in moral and legal contexts are further, additional interests; we can call the process by which these interests arise from context "pragmatic" either because they are somewhat irrelevant to the basic mechanism of selection, or else — more positively — because we are gesturing towards some general account of how context determines certain interests, and thus ultimately plays a role in causal selection.

These are merely sketch responses on behalf of the contrastive approach, and it is hard to assess their merit without a fuller account. Nevertheless, there are weaknesses in the general strategy of appealing to pragmatics.

First, it is worth noting that a positive account of the pragmatics of causal selection is called for. Merely alluding to pragmatic considerations, as a sort

of "black box", will not provide a satisfactory theory of selection. It is easy to invoke pragmatics in an entirely platitudinous way, amounting to nothing more than the admission that selection is flexible and depends on context. The question is *how*. Until that question receives an answer, the two lines of objection cited are not satisfactorily rebutted. Merely repeating that selection is flexible and depends on context will provide only the weakest, most general explanation of why we select what we do. Likewise, it is little advance simply to assert that the choice of contrast is not entirely "up to us" in moral and legal contexts: to justify the weight we put on selection in those contexts, we need to say something more about what *does* determine choice of contrast.

Second, there is no very obvious candidate for a pragmatic theory answering these needs. When devising a pragmatic account, obvious places to look for inspiration include Searle's theory of speech acts (Searle 1969, 1975) and Grice's theory of conversational implicature (Grice 1975, 1978). It is hard to see how a theory of speech acts would be directly relevant to our present problem. Grice's work might be more directly useful, concerned as it is with the distinction between the strictly true and the true-but-misleading. To say that the presence of oxygen caused the flame might be false but misleading; and indeed this Gricean picture is the one Lewis suggests (Lewis 2004a, 101). This general remark, however, does not amount to an account of the pragmatics of causal selection. It is far from obvious how the Gricean maxims are to be applied to generate such an account (Menzies 2004, 147), and some of the notions appearing in the maxims, such as the notion of relevance, might seem uncomfortably similar to what we had hoped to analyse.

Switching our attention to the objection from justifying moral and legal practices, we might pretheoretically suppose that the fact about who caused the warehouse to burn is to be settled by principles which are not moral or legal. It is a factual question, we might think. A pragmatic account of contrast choice might violate this pretheoretic supposition, if the principles governing selection in moral and legal contexts differed from those in other contexts. This would be the most natural way for a pragmatist to account for the flexibility of selection. But it would be problematic. We might, on reflection, admit that the moral or legal context determines a different selection from, say, a scientific explanatory context. But we do not, I think, normally suppose that there are special moral or legal *principles* of selection. We admit that causal selection is flexible, and that our interests are determining factors; but it does not follow that any special principles apply in moral or legal contexts. If they did, the justificatory

problem would become pressing, because moral and legal evaluation of facts is naively supposed to take place after, and independently of, those facts. But on such a picture, those facts would be determined by principles which were specific to the moral and legal context. This is a particularly strong kind of context-dependence: the kind that we normally accept, I suggest, implies only that context plays a determining role in causal selection, having some sort of input — not that it also supplies the principles governing that determination.

These criticisms of the pragmatic/contrastive account need to be stated with care. The initial criticism of the contrastive account is that, without an account of contrast choice, it provides an incomplete account of causal selection. The need for an account of contrast choice is not felt by many (but see Menzies 2004). Presumably this is because the pragmatic considerations that govern it, however difficult to state, are thought to be neither especially problematic nor especially interesting. My main argument against such a view is to insist that it leaves our account incomplete. There are two ways I want to avoid being misunderstood. First, I do not want to be taken as questioning the flexibility of selection. The suggestion that selection includes a pragmatic element should be sharply distinguished from the suggestion that it is flexible, meaning that we will select different things in different contexts. The latter is an evident fact. If the appeal to pragmatics is taken as a mere restatement of this fact, as I think it sometimes is, then obviously it is no analysis. If, on the other hand, it is taken as an analysis of the flexibility of selection, then it is inadequate because no theory has been provided. Second, I do not want to be taken as saying that there can be no pragmatic account of contrast choice. Rather, I am suggesting that no account has been given. Lewis gestures at an account in terms of conversational implicature, but he does not provide one. At this stage, I do not want to claim anything stronger than that no adequate account of contrast choice has been provided, and thus that the contrastive account of selection is incomplete. Once I have proposed my own account, I will be in a position to make the stronger claim, that the pragmatic/contrastive account is explanatorily suboptimal with respect to the objections we have considered.

4.4 The Reverse Counterfactual

4.4.1 Contrast Choice and Counterfactuals

Suppose we accept that the mechanism of causal selection is contrastive, and ask: What determines the choice of contrast? I strike a match, and it lights. On the contrastive account, the reason we select the strike as the cause of the flame is that we entertain a contrast between this situation, in which the match lights, and certain others, in which it does not light, but in which oxygen is still present. The question, then, is why we pick this contrast rather than a contrast where the match is still struck but the oxygen is absent. The strength of the contrastive approach is that it allows that we sometimes *do* pick such a contrast; the weakness is that it gives no indication of how or why.

At this point, instead of crying "pragmatics", let us remember Menzies' point about the concept of difference-making. Causes are widely considered to make a difference to their effects: on Lewis's account, the difference they make by occurring is between the effect occurring and not. But when we ask which, among these events, is *the* cause, we might be understood as asking what makes, not just *a* difference, but *the* difference. Then the question is: *The* difference between what and what? One obvious answer might be: between the effect occurring, and the counterfactual scenario in which the effect does not occur. Contrast thus specified, the rest of the process occurs as described by the contrastive account. The difference between the case where the effect occurs, and the contrast case where it does not, is selected as *the* cause. There may be more differences than we mention, so an appeal to pragmatics may be ineliminable; but we have significantly confined its role.

The idea, in short, is that the contrast case is often just the closest possible world in which the effect does not occur. I shall propose an account based on this idea, and argue that it answers the objections raised against the contrastive/pragmatic strategy.

4.4.2 The Difference Between Cause and Condition

I now propose a counterfactual which is true of causes and not true of mere conditions. The counterfactual is arrived at by reversing the Lewisian counterfactual, to yield:

> **The Reverse Counterfactual Necessary Condition on Causation**

If c causes e then $\sim E > \sim C$.

Call the counterfactual in this necessary condition the *Reverse Counterfactual*. Clearly the Reverse Counterfactual will often be a backtracking counterfactual — one whose antecedent denotes something temporally later than the consequent. Lewis and others have given backtrackers a bad name, but I view them more favourably, for reasons indicated in Chapters 2 and 3.

I strike a match in a warm, dry room, and it lights. I assert that if the match hadn't lit, then I wouldn't have struck it. In support, suppose you come into the room to find me holding a lighted match. Now ask what you would have thought, had the match not been lit. Unless you have reason to think otherwise, you would assume I had not struck the match. To dispute this is to question the reliability of matches quite generally, any time they light; a perverse scepticism, belied by their usefulness. (This is an application of the Inference Test.)

Immediately it will be objected that, if the match hadn't lit, perhaps I would still have struck it, but clumsily, without sufficient force or speed. After all, matches don't always light. I think this is not an objection, but indicates an advantage of the account. For what we must say to get round the apparent objection is that, from an unlit match, we would infer that the match was not struck *well* — not hard enough, not fast enough, or some such. These factors are indeed part of the cause of the flame: they are causally relevant. It is no harm if the Reverse Counterfactual identifies them as such. For when we say the match was not struck well, we are not introducing an endless *ceteris paribus* clause, including all the necessary conditions for the flame, as we shall now see.

The Reverse Counterfactual sharply differentiates the striking of the match from the presence of oxygen. There is no reason to suppose that if the match had not lit, there would have been no oxygen in the room. In this particular case of a match lighting, in this warm, dry, airy room, we do not infer the absence of oxygen from an unlit match. So we cannot use that manner of argument to support the claim that, if the match had not lit, the room would have suddenly evacuated. Nor does this counterfactual have any intuitive plausibility.

The contrast can be brought out sharply in worlds-talk. It is, at least, arguable that the nearest worlds[5] where there is no flame are worlds where

[5]Here and at other points I make the Limit Assumption. This is purely for the sake of exposition. The Limit Assumption plays no role in the account.

there is no match-strike. Whereas it is not plausible that the nearest worlds where there is no flame are worlds where there is no oxygen.

4.4.3 Flexibility

Perhaps the most obvious question we might ask about this proposal is whether it can account for the flexibility of selection. It has been heavily emphasised that what we select depends on context, both of inquiry and of occurrence. To reflect this flexibility, the Reverse Counterfactual must be true of different events in different contexts. It must explain why we are sometimes willing to say that the presence of oxygen caused the flame rather than the match-strike, and why we sometimes unselectively allow that they are both causes.

The Reverse Counterfactual, like any other counterfactual, depends both on context of inquiry and on context of occurrence. As a result it can explain the flexibility of causal selection relative to both kinds of context. In 3.1 we saw that counterfactuals are relative to the context of inquiry: depending on the conversation, we change our views about whether Caesar would have used the atom bomb or catapults had he been in charge in Korea.

But this is not relativity to context of occurrence: what changes our views of the counterfactuals in that case is a shift of conversational context, not a difference between the context in which some actual event occurs and that in which some other actual event occurs. I suggest we understand relativity of a counterfactual to context of occurrence as follows. A counterfactual $\sim A >\sim C$ is *relative to context of occurrence* just in case:

- the counterfactual $\sim A >\sim C$ has a different truth value to $\sim A' >\sim C'$, where
- a and a' are actual events of the same kind, as are c and c'.

We need only a very weak notion of "same kind": the events need not be exactly similar; rather, they must be similar enough, such that the reason for giving different truth-values to the two counterfactuals is not that the events concerned are different. I suggest that counterfactuals are relative to context of occurrence, in the sense given. The context in which the events they mention occur is a determining factor of the truth-value of the counterfactual.

The point is obvious when we consider an example:

> If I hadn't pressed the light switch, the light wouldn't have lit.

Consider this counterfactual in relation to two events. First, I am on my own at home and I switch the light on as dusk approaches. In that case, it seems that if I hadn't pressed the light switch, the light wouldn't have lit.[6] Second, suppose my wife is at home too. In that case, we might suppose that if I hadn't pressed the light switch, she might have done; so it is false that, if I hadn't pressed the light switch, the light wouldn't have come on. This counterfactual, then, is true of one occurrence when I am home alone, and false of a similar occurrence when I have company. The point is also obvious on Lewis's semantics. If I am on my own in the house, the closest worlds where I don't press the switch are ones where the light stays off; whereas if my wife is at home, at least some of them are worlds where she switches the lights on. It seems, then, that counterfactuals are relative to the context of occurrence. Hence the Reverse Counterfactual account makes causal selection relative to the context of occurrence.

Since it is a counterfactual, therefore, the Reverse Counterfactual depends on both context of inquiry and on context of occurrence. Now the question is: Does it do so in the same way as causal selection? First, consider this same match-strike, in a warm dry room, from the perspective of a chemistry lesson. In that case, we might well also pick the oxygen as a cause. I suggest that we explain this by saying that the context of that chemistry lesson determines a resolution of vagueness on which the closest no-flame worlds include worlds without oxygen (though perhaps to complete this story we need an account of joint causes, given in 4.5.2 below). Second, consider a match-strike suitably similar to this one, but occurring in a different context: a chamber that is ordinarily oxygen-free (perhaps as part of some manufacturing process). On this occasion, oxygen has leaked in, so when the match lights we cite the oxygen as the cause of the flame (cf. Hart and Honore 1985, 10). I suggest that this, too, is reflected by the Reverse Counterfactual. We select the oxygen in this situation because the nearest worlds where there is no flame are ones where the match is still struck, but the oxygen is absent as usual.[7]

[6]Of course, it is *possible* that the light could have come on some other way, but that is the stuff of thrillers and ghost stories — thrillers if someone else turns out to be present after all, ghost stories otherwise. This point just emphasises the way context of occurrence affects the truth of the counterfactual.

[7]It might be objected that, in citing the Reverse Counterfactual as an explanation of our selective practices, I am relying on an implied claim that the Reverse Counterfactual is sufficient for causation (whereas officially, I maintain that it is necessary but not sufficient). I owe this objection to Arif Ahmed. The correct reply, I think, is that the objection artificially restricts the explanans. Spelled out, my claim is not that the Reverse Counterfactual explains causal selection: rather, it is that the *fact that the Reverse Counterfactual is a*

It seems, then, that the Reverse Counterfactual can provide an account of the flexibility of selection. The reason is that counterfactuals depend on context. So my suggestion amounts to this: that the principles governing the context-dependence of causal selection are just those which govern the context-dependence of counterfactuals. This is not quite the whole story about flexibility, and we will discuss it further in 4.6.1. But before we do that, let us consider how an account of this shape handles the objections we considered against the contrastive account (4.4.4), and then flesh out the details of the proposal (4.5).

4.4.4 Two Objections Answered

If this is right, then we have a distinction between cause and condition: we can only show the Reverse Counterfactual to be true of the cause, and not of mere conditions. Does this account of selection handle the objections raised in 4.3.2? Both those objections pushed the point that, if selection is to be accounted for contrastively, then an account of contrast choice needs to be included, and in 4.3.3 I argued that appeals to pragmatics were neither adequate in their current form, nor particularly promising. The Reverse Counterfactual is equivalent to the contrastive strategy plus the claim that the contrast is just the contextually appropriate counterfactual situation (nearest possible world). As such, it provides an answer to the general problem which those objections pushed, because it implies an account of contrast choice. That is the general answer: let us see how the proposal answers the two specific lines of objection which were intended to push that point.

The Reverse Counterfactual answers the objection that the contrastive account is insufficiently explanatory. Menzies noted that the notion of difference-making is employed twice, on the contrastive account: "first in the analysis of 'objective cause' as something that makes a counterfactual difference; and then again in the contrastive explanation account of the 'context-sensitive cause"' (Menzies 2004, 150). The proposed account achieves greater theoretical unity,

necessary condition on causation explains causal selection. In offering my analysis as an explanation, I may indeed be relying on a claim that a certain fact (namely, the truth of my analysis) in some sense suffices for causation to be selective. But that reliance does not amount to (and in fact is incompatible with) suggesting that the Reverse Counterfactual itself is sufficient for causal selection. By way of analogy, the claim that counterfactual reliability is necessary for knowledge might be advanced as an explanation for the fact that we tend not to regard counterfactually insensitive beliefs as knowledge (if indeed that is a fact). Clearly, this would not imply that counterfactual sensitivity of belief is also sufficient for knowledge.

however, since it applies the notion of difference-making in only the counterfactual way. Causes make the difference to their effects; I have suggested that causes make the difference by *being* the difference between the actual case where the effect occurs and the counterfactual case where it does not. (Which case this will be depends on context, of course, as discussed in 4.4.3.) Relatedly, this feature answers the objection to the contrastive account that it makes selection doubly dependent on context: first for resolving the vagueness of the Lewisian-causal counterfactuals, and then in determining the selection among those counterfactuals. Context-dependent selection *between* true counterfactuals should not be confused with context-dependence *of* counterfactuals. The Reverse Counterfactual considerably reduces the role of former in accounting for causal selection.

I suggested that the prediction/explanation symmetry could be employed to bring out the explanatory shortcomings of the contrastive account. It can equally be employed to bring out the explanatory successes of the Reverse Counterfactual. The Reverse Counterfactual yields concrete predictions about the events we will select as the cause in a given context. We saw this just now: the argument that the Reverse Counterfactual is true of the match strike and false of the presence of oxygen amounts to a theoretical prediction. If cases can be found where the Reverse Counterfactual appears to be true of a mere condition, or false of the event we select as the cause, then we will need to revise or abandon the approach. I take it that this falsifiability is a further point in favour of the proposal over the contrastive account; it is hard to see how a contrastive account could be vulnerable to such counterexamples while the pragmatic account of contrast choice remains schematic.

The objection from moral and legal justification is also well-handled by the Reverse Counterfactual. Our arsonist might object that the builder is just as much a cause of the fire, and thus that we are treating her unfairly in sending her to prison. The first way in which the Reverse Counterfactual offers a superior justification for imprisoning the arsonist is simply that it offers a better explanation of this particular selection. Since there is a naive supposition that justice is not arbitrary, justification will be served by explanation. I have just argued that the Reverse Counterfactual is explanatorily superior to the contrastive account. We can reply to the arsonist that the builder is not the cause of the fire: given the circumsances, it was the arsonist, and not the builder, who made the difference between the fire and no fire.

To contest this point, the arsonist would need to argue that the builder

made as much of a difference. This argument is of the same kind as the debate which might arise about whether Caesar would have used catapults or the atom bomb in Korea. Debates of this sort are not confined to the moral or legal context. We have, I suggest, a considerably improved justification for sending the arsonist to prison if we can reduce the debate to a kind which is not specific to the moral or legal sphere. For in those spheres, there is a naive requirement that fact and evaluation should be distinct. We do not, in moral and legal contexts, think that the arsonist caused the fire for moral or legal purposes. We think she caused the fire, independent of any moral and legal evaluation: and that is *why* we condemn her morally and legally.

The Reverse Counterfactual does not eliminate context-dependence of selection, but it offers a unified account of the way in which selection depends on context. There is no special *kind* of consideration which comes into play in moral or legal contexts; there are no special moral or legal principles of selection. It is just the same kind of consideration that comes into play in other, non-moral and non-legal contexts, such as the context of lighting a match in a warm, dry room. I suggest that this considerably helps to justify selection in moral and legal contexts. Selection might still depend on the moral or legal context, but that is nothing special about selection for moral and legal purposes: selection always depends on context, whatever kind of context it is.

4.5 Refinements

4.5.1 Absent Causes

We have defended the Reverse Counterfactual with regard to a couple of examples. Granting that the initial defence is successful, we should ask whether results obtained by considering these examples are likely to generalise. One obvious potentially non-general feature of the examples is that they are somethings, not nothings. If we wish to allow causation by absences, we should check the Reverse Counterfactual against an example of causation by absence. Suppose I fail to water the plant on my window-sill, and it wilts. Normally we would say that it was my failure, not yours, nor the Queen's, which caused the plant to wilt. The Reverse Counterfactual agrees. If the plant had not wilted, I would have watered it; but it is false that if the plant had not wilted, you would have watered it — after all, you probably don't even know where I live. And it seems at least as far-fetched that Her Majesty would have paid a visit. We can support the Reverse Counterfactual, and undermine the others, by the

Inference Test. If, in these circumstances, my plant had not wilted, then (if you somehow came to hear of it) you would infer that I had been watering the plant, and you would neither infer that the Queen, nor that you yourself, had watered it.[8]

It is sometimes thought that allowing absences as causes opens up an enormous can of worms, especially on Lewis's view: for in any given case there is a truly huge number of absences which, had they been presences, would have prevented the effect. This motivates Beebee, for example, to deny that absences cause.

> There is no genuine causal difference between those cases that common sense judges to be cases of causation by absence and those that it judges not to be cases of causation by absence.
>
> (Beebee 2004, 293)

Beebee offers a critique of various solutions to the question of how we select the putatively causal absence from among a series of increasingly outlandish absences which, had they been presences, would have prevented the effect. For example:

> One of the causes of your reading these words right now is the absence of a lion from the room.
>
> (Beebee 2004, 298)

As far as Beebee can tell, efforts to prevent results like this all fail.[9] The absurdity of a view which licenses assertions like the one above is a significant factor driving Beebee to adopt the opposite view, that no absences cause.

There may be reasons to deny that absences cause (cf. Lewis 2004b, Menzies 2002), but it should be pointed out that the problem of selection is not one of them. For then parity of reasoning would dictate that we deny that presences cause, since they too suffer from the problem of selection. Perhaps the problem is not as severe, in the sense that there are fewer positive mere conditions (consisting of actual events or states of affairs) than there are negative ones (consisting of the absence of possible states of affairs). But it is far from clear that the severity of the problem is directly proportional to the number of conditions. One of the things which, I argued, makes selection difficult

[8] I am indebted to Kit Patrick for first alerting me to the possibility that the Reverse Counterfactual might help with selection among causal absences [Patrick, 2005].

[9] The efforts she criticises do not bear much resemblance to mine: she focuses on the special event strategy of Hart and Honore.

to explain is the moral and legal weight we put on it. I argued that point by comparing the imprisonment of an arsonist with just one alternative, the imprisonment of the builder. It is not immediately clear that explaining a choice among more alternatives is any harder than explaining this choice between just two. In any case, selecting the cause among positive mere conditions is often of great importance; it is not clear that ruling out mere conditions that are absences is more important.

Moreover, at least one solution — the Reverse Counterfactual — to the problem of selection among presences will generalise to absences, as I have just argued. There may be good independent, perhaps theoretical reasons to deny that absences cause, but our present concern is to map our ordinary concept of causation with counterfactuals, and in this context absences seem as apt as presences to count as causes. There is no special problem of selection for causal absences, and no special solution either.

4.5.2 Joint Causes

Another notable feature of the examples we have been working with is that they are not cases of joint causation. In joint causation, the Reverse Counterfactual is false of each of the jointly-causal events. Suppose you and I together lift a table. Assuming we are both similarly motivated (or similarly lazy), and assuming we are both up to the task, then there is no particular reason to say that, if the table had not risen, you would not have lifted: for you might have done, and I might not have. The same goes for me. The Reverse Counterfactual is therefore false of both our efforts; yet we presumably consider ourselves to have together caused the table to lift. We distinguish our efforts from the continued solidity of the ground, which is a mere condition of our accomplishment in this case.[10]

The obvious solution is to say that if the table hadn't risen, then at least one of us would not have lifted, and thus that $\sim E >\sim (C_1 \& ... C_n)$ where $c_1, ..., c_n$ jointly cause e. Unfortunately, however, if we allow that counterfactuals with conjunctions for may satisfy the necessary condition on causation imposed by the Reverse Counterfactual, we have an immediate problem. For any propositions P and Q, $\sim (P \& Q)$ follows from $\sim P$. So any conjunction will qualify as the cause of any effect, provided the real cause is one of the conjuncts. If the match had not lit, then it would not be the case that I struck the match and you scratched your ear. But my strike causes the flame without

[10] In the following solution I am indebted to Torben Rees.

any assistance from your scratch.

To block this difficulty we could seek to disallow counterfactuals with conjunctive consequents. But that would rule out the proposed account of joint causation. Therefore I suggest we further stipulate that each conjunct must figure ineliminably. So if $c_1, ..., c_n$ are joint causes of e then (i) $\sim E >\sim (C_1\&...C_n)$ and (ii) for every non-empty proper subset $\{c_x, ..., c_y\}$ of $\{c_1, ..., c_n\}$, $\sim (\sim E >\sim (C_x\&...C_y))$. That is, there must be no proper subset of the candidate joint cause which itself meets the condition initially proposed for qualifying as the joint cause. If there is, then the others are eliminable.

For example, you help me lift a table: we jointly cause it to rise. If the table had not risen, then it is not the case that both you and I would have lifted. However we cannot, with confidence, say that you would not have lifted, nor that I would not have lifted. Thus neither of our lifts qualifies on its own (so neither is eliminable). This appears reasonable and intuitive. Moreover it is supported by the reasoning which led us to conclude that the Reverse Counterfactual failed for joint causes: for that turned on noticing that we could not say that if the table hadn't risen, you wouldn't have lifted, because if the table hadn't risen then you might have, but I might not have, and vice versa. Finally, note that the solidity of the ground fails the condition. In the circumstances we would surely agree it is false that, if the table had not lifted, the ground might have given way. Otherwise the menial task would acquire a new urgency.

4.5.3 Mere Conditions

We have given an account of the difference between cause and mere condition. But we have given no account of the difference between a mere condition and a causally irrelevant event. Endorsing a selective notion of causation does not mean denying that the presence of oxygen is relevant to the lighting of the match in a way that the football results fail to be. I suggest that a condition d_1 for an event e is just a counterfactually necessary condition: that is, an event d_1 satisfying the Lewisian counterfactual, $\sim D_1 >\sim E$. A *mere* condition can then be defined as a condition which fails to satisfy the Reverse Counterfactual: that is, a condition which is not the cause.

One consequence of this view is that the cause c of an event e might not be a condition for e at all. For it does not follow from $\sim E >\sim C$ that $\sim C >\sim E$. This feature of the account will be exploited in Chapter 6 to provide a solution to preemption counterexamples, and overdetermination problems

more generally, which arise directly from the fact that causes are not always counterfactually necessary conditions for their effects.

We might wonder, however, whether this view of the cause/condition difference is too extreme. What, if any, is the connection between the two notions? Intuitively, a mere condition is still *causal*. On my account, *the* cause is given by the Reverse Counterfactual, which is false of mere conditions; what, then, makes mere conditions causal? Another way to push the same point would be to ask about the notion of causal relevance. Presumably both causes and mere conditions are causally relevant to their effect, yet I am suggesting entirely distinct criteria for causehood and mere conditionhood. What, then, is the nature of the causal relevance which — intuitively — they share?

The obvious answer is the disjunctive one: c is causally relevant to e just in case c is a condition or a cause of e. Except that they both satisfy a counterfactual, they have nothing more in common. I do not think this is a problem. Lewis held up the prospect of analysing all of causation with a single counterfactual, but given the complexity of causal concepts and the range of uses to which we put them, we should perhaps not be surprised if a counterfactual analysis of causation must appeal to more than one counterfactual. Moreover, the fact that Lewis's account of causation is so radically unselective strongly suggests that he achieved such a unified analysis at the expense of part of what he was analysing.[11]

A further point can be adduced in support of the idea that being a cause is something different from being a condition (mere or otherwise). The Reverse Counterfactual might be thought of as making causes counterfactually sufficient for their effects, in a certain sense (to be clarified in 6.1). In that case we can explain why causes are often also conditions. If c is counterfactually sufficient for e, and if there is nothing else that is counterfactually sufficient for e, then c will be counterfactually necessary for e (since, if nothing else will suffice for e, then without c, e would not occur). Thus causes will also be conditions when there is no kind of causal redundancy. This would explain why, by our lights, Lewis mistook conditions for causes. It would also explain why

[11] There is obviously a great deal more to say about causal relevance. Kvart suggests that we should employ a notion of causal relevance in analysing counterfactuals (cf. Kvart 1986, 1991, 1994). Consequently, analysing causal relevance in terms of counterfactuals "put[s] the cart before the horse" (Kvart 1994, 98). Nevertheless, I do not think our views need be incompatible. All I claim here is that causal relevance has counterfactual entailments — that is, that counterfactual dependence of some sort or another is a necessary condition on causal relevance. This is consistent both with analysing counterfactuals in terms of causal relevance and with my defence of the Reverse Counterfactual as a necessary condition for causation (see Chapter 7).

his analysis struggles with redundancy of various kinds. For c can be the cause of e even when it fails to be a condition, and cases of redundancy are exactly cases where c causes e without being a condition for e. This must count in favour of my contention that causes need not be among the conditions for an effect. It will also provide the starting point for our discussion of redundancy in Chapter 6.

4.5.4 A Principled Argument

The Reverse Counterfactual has been defended and developed with respect to a particular example, but I have also stressed its connection with the quite general idea that causes make the difference to their effects. Now that I have proposed my account, let us revisit that idea.

The underlying thought is that it is implausible to maintain that some event caused another without maintaining that its occurrence *made the difference* between the occurrence of the effect and its failure to occur. That thought is widely accepted,[12] and on its own it does not presuppose a selective notion of causation: Lewis's account can be seen as specifying that an event c makes the difference to e just in case, if c hadn't happened, e wouldn't have happened. It follows that a lot of events make the difference to any given effect.

I have suggested a different understanding of difference-making, however: that an event makes the difference by *being* the difference between the actual case and the counterfactual scenario where the effect would not occur. This view seems, to me, implicit to some degree in various other authors: in Mill's Method of Difference [Mill, 1887], in Mackie's remarks about causes and effects being differences relative to a causal field (Mackie 1974), in Lipton's Difference Condition on contrastive causal explanation [Lipton, 2004], and even in Lewis's account of contrastive causal explanation [Lewis, 1986b]. But if the view I propose is correct — if an event makes the difference by *being* the difference between the case where the effect occurs and the counterfactual scenario where it does not — then the Reverse Counterfactual must, in principle, be true of all causes. For it is just the counterfactual supposition described — supposing what would be the case if the effect hadn't occurred.

"The" counterfactual supposition is of course a misnomer, because counterfactuals are vague; in a given case, the appropriate counterfactual scenario (possible world) for determining the truth of the counterfactual will depend on

[12]For example: Mill 1887, Mackie 1974, Lewis 1973a, Hart and Honore 1985, Lipton 1993, Menzies 2004, Schaffer 2005.

context of inquiry and of occurrence, as we have discussed, in just the way that the vagueness of counterfactuals is always resolved. I have already indicated how this allows us to account for the flexibility of causal selection. Now that I have filled in some of the details of the account, let us consider some more general questions about it. We shall consider the relation between the Reverse Counterfactual and contrastive explanation (4.6.1). This discussion will be useful in rebutting an objection, due to Peter Lipton, that the Reverse Counterfactual cannot match the flexibility of our causal selective practices (4.6.2). Then we will briefly compare the Reverse Counterfactual with a proposal of Peter Menzies' — another account seeking to assimilate the context-sensitivity of selection to the context-sensitivity of counterfactuals (4.6.3).

4.6 Objections and Comparisons

4.6.1 The Reverse Counterfactual and Contrast

Despite constituting a rejection of the strategy of assimilating causal selection to contrastive causal explanation, the Reverse Counterfactual account of selection shares the key idea behind the contrastive approach: that the cause is what "makes the difference". From this fact, I hope to draw some support for the Reverse Counterfactual account of selection, and to explain both the successes and the shortcomings of the attempted extension of the contrastive mechanism from causal explanation to causal selection.

Although I am advocating a radically selective account of causation, this is entirely compatible with citing mere conditions in causal explanation. The beaker of water boiled because I lit a bunsen burner under it. Nevertheless I might explain the boiling of the water in terms of the molecular structure of water and the atmospheric pressure, because the chemistry lesson makes that appropriate. I already suggested that the Reverse Counterfactual is true of the atmospheric pressure in such a situation, because of the way it depends upon context. There is, however, another point to be made. Sometimes, explanation and causation may come apart. The cause is not always the explanation. The teacher will explain a flame by appealing to the presence of oxygen. Yet even chemistry teachers will not generally accept the presence of oxygen to have caused the small bonfire at the end of the bench.

The Reverse Counterfactual is, as we have seen, equivalent to the contrastive account of selection plus a counterfactual account of contrast choice. This means that it fits very neatly with a contrastive account of causal ex-

planation. Normally, on the view I am proposing, the choice of contrast is determined by the nearest possible worlds, and which worlds are nearest depends on context in the ways we have discussed. When we explain, however, we might be seen as manipulating the standards governing which worlds are nearest. One way to do this is by specifying a contrast. For example, asking "Why did the match light?" naturally yields the answer, "Because it was struck". If, however, we ask, "Why did the match light, rather than not lighting, when it was struck?", we have, as it were, manually adjusted the context so as to ignore the most obvious counterfactual scenario where the match is not lit. The striking of the match has been specified as occurring in the relevant antecedent-worlds. Now we are asked to look for the nearest counterfactual scenario where the match does not light, but was still struck. Assuming, as I think is legitimate, that we hold constant other features of the ordinary match-strike — the dryness of the match, the competence of the striker — we arrive at worlds where there is no oxygen.

If this is correct, then we have shown how the contrastive account of explanation fits with the proposed account of causal selection. And more excitingly, we have found an answer to one of the questions we started with. For we can now see *why* causation is used as a selector in explanation. We use causation to select in explanation because causation is selective. Contrastive causal explanation is a way we have of extending the selective aspect of causation to identify events which are not the cause, but which are causally relevant.

4.6.2 When the Cause is Counterfactually Stable

We might be suspicious of the suggestion that the cause is always the most counterfactually "fragile" event — the one that is absent in the nearest worlds where the effect is absent.[13] Might there not be examples where the event we call the cause is not, after all, the event which is absent in the counterfactual scenario? Suppose we are attempting to measure background radiation. We achieve a reading, and say that the cause of the reading was background radiation. However, if we had achieved a wild reading or no reading at all, we would have inferred some problem with the equipment, and certainly not that the background radiation was markedly different from what was previously thought.[14]

[13]This sense of "fragile" is meant to aid understanding, and is not the Lewisian technical sense, on which an event's fragility is a measure of the richness of its essence (Lewis 1986a, 196).

[14]The objection and counterexample I owe to Peter Lipton.

Scientific experiments are often like this, and put pressure on the manner of arguing from inference which I have employed, since they are contexts where what we infer may vary depending on the outcome of the experiment. A good reading would be taken to confirm the presence of background radiation, but a bad one would lead us to infer that the equipment is not working. In addition, we might wonder what sort of semantics of counterfactuals could ever yield the result that a world where background radiation is absent is closer than one where a delicate piece of equipment malfunctions.

In response, I suggest that there is a quite ordinary sense, in which the good reading may naturally be said to be caused by correctly functioning equipment, just as the bad reading may be said to be caused by poorly functioning equipment. That is why the scientists who first measured background radiation were so proud: they had, in a very obvious sense, caused the readings. They made the difference between readings and no readings. But the background radiation did not: it would have been present, surely, regardless of whether it had ever been successfully measured. On the other hand, if we assume the equipment is functioning correctly, then the background radiation *does* make the difference between reading and no reading. In the context in which we say that the radiation caused the reading, we hold constant the working of the equipment and a range of other local conditions — even *per imposibile*, and even if we don't know what they all are. Then the radiation makes the difference between reading and no reading; and then, if there had not been a reading, there would have been no radiation.

This example serves further to clarify the relation between mere conditions and causes. What is a mere condition in one context of inquiry may be a cause in another. I have suggested that the context-sensitivity of the Reverse Counterfactual reflects this fact. The example also shows how contrastive explanation extends the selective aspect of causation. We do not need to specify or even know about all the factors which determine the truth value of the Reverse Counterfactual in the context of this experiment. But one obvious way we might manipulate those factors, communicate how we understand them and direct the interests of others, is by specifying contrasts. Thus I might contrast the reading with the lack of a reading on the same properly functioning piece of equipment, in order to direct your attention away from the possibility of mechanical error and towards the phenomenon we are seeking to measure.

Nor is it an accident, I think, that this counterexample should be alleged in an experimental context, where explanation is a prominent interest. The

claim that the radiation causes the reading is arguably a causal explanation in disguise, rather than denoting the particular cause of this particular event. The experimental context is a highly developed and consciously refined context of inquiry, deliberately constructed to enable us to isolate what would, normally, be mere conditions. In the experimental context, it may be the case that if the reading had been different, then the radiation would have been different. Contrasts are a simple and powerful way to manually adjust the context-dependence of counterfactuals.[15]

Thus asserting that the Reverse Counterfactual is true of causes might also help account for the success of the Difference Condition on contrastive causal explanation. That condition requires that a good explanation be a relevant causal difference between the explanandum and some contrast. When the contrast is simply the nearest counterfactual situation where the explanandum does not occur, as in many salient cases it will be, the Reverse Counterfactual picks out as the cause an event which the Difference Condition picks out as a good explanation. The contrastive account of selection sees causal selection as an extension of the selective mechanism employed by explanation of a certain sort. The approach I am advocating sees the situation the other way round: the contrastive account of explanation extends the selective aspect of causation, to cover cases where our explanatory interest is not in the cause, but in a mere condition. It effectively generalises the selective mechanism of our causal concept. The selective aspect of causation is part of the reason why it is a useful explanatory tool.

4.6.3 Menzies' Causal Models

My account of causal selection seeks to account for the flexibility of selection by assimilating it to the context-sensitivity of counterfactuals. It is widely accepted that counterfactuals are sensitive to the context of inquiry, and I have argued that they are also sensitive to the context of occurrence — a point which I think is fairly obvious. Menzies takes causal selection seriously, and rejects

[15]Lipton discusses the possibility that contrasts might be a special way of resolving the vagueness of counterfactuals so as to allow backtrackers (Lipton 1993). He notes that if something more substantive could be said about how contrasts resolve vagueness to allow backtrackers, his account of contrastive explanation could employ backtrackers and thus would move closer to Lewis's. I have argued that vagueness is not at issue, and that we can endorse backtrackers directly, in standard contexts. The place I envisage for contrast is as helping specify which antecedent-worlds we are interested in — a particularly useful tool with which we can cut through a lot of the inherent vagueness of counterfactuals, whether foretracking or backtracking.

the contrastive strategy, as we have seen. He also seeks to account for causal selection "by giving an account of difference-making in terms of context-sensitive counterfactuals" (Menzies 2004, 178). His idea is to make causal judgements relative to *causal models*, which are a technical elaboration of Mackie's idea of a causal field. Causal models determine their own similarity-ranking among possible worlds, Menzies suggests. His discussion is sophisticated and I shall not expound the details here, but the nub of it is that to be a cause, an event c must satisfy both $\sim C >\sim E$ and $C > E$ (Menzies 2004, 170–1). Mere conditions will satisfy the former condition, but fail the latter, if we drop the Centering Assumption, which Menzies says we should (Menzies 2004, 166). Then it is likely false that in all the nearest possible worlds (relative to a given causal model) where the oxygen is present, the match lights.

Apart from its complexity, which I have suppressed in my thumbnail sketch, this solution has two serious drawbacks. The first is that rejecting the Centering Assumption is a serious business. I have argued (Chapter 3) that endorsing backtrackers does not constitute a serious departure from Lewis's semantic theory — the asymmetry of counterfactual dependence is a further thesis, which can be dropped or modified. Dropping the Centering Assumption, on the other hand, means either abandoning or supplementing the idea that counterfactuals are to be analysed in terms of comparative similarity between worlds. Although philosophers are rather sanguine about dropping the Centering Assumption when it suits them (as I mentioned in 2.2.2), it bears emphasis that no analysis of counterfactuals in terms of non-comparative similarity between worlds has gained currency. Until such an analysis is advanced, any account which rejects the Centering Assumption incurs serious losses, and must score highly in other respects if it is to be preferred to an account which is compatible with the Centering Assumption.

Second, the causal model account does not seem to perform especially well against the objections we considered against the contrastive account. It is hard to see how it generates any real predictions. We are free to stipulate that, when I strike a match in a warm dry room, the causal model consists in a certain set of worlds, such that the match-strike is a difference among them whereas the oxygen is not. But it is hard to see how this amounts to a prediction: without a substantive theory of non-comparative similarity, we can say what we want about which worlds are closer, and we will be almost immune to challenge. I have suggested that this indicates a lack of explanatory power. Likewise, the suggestion that selection is relative to a causal model does

not fully explain the moral and legal weight we put on it, unless we can say something further about what gives rise to the causal model. Menzies focuses on the way causal models generate spheres of "normal worlds" against which differences are made (Menzies 2004, 160–69), and does not emphasise any particular problem with the origin of causal models. But my arguments about the generation of contrasts apply equally to causal models: in accounting for ordinary selective practices, it is essential that we understand, not merely how causal models (or contrasts) give rise to selections, but what gives rise to causal models (or contrasts) in the first place. Menzies denies that he is committed to a "crude relativism" on which causation is mind-dependent, and he denies that all causal models are as good as each other because of the existence of natural kinds: but he accepts that "a plausible metaphysics is likely to allow that any particular spatiotemporal region instantiates several kinds of systems" (Menzies 2004, 159), and causal models are defined over systems (Menzies 2004, 160). In his favour, Menzies seeks to employ the context-sensitivity of counterfactuals to analyse the context-sensitivity of selection — a strategy which I also endorse. So it may be that Menzies would have the resources to say more about how causal models — and hence causes — depend on the context of occurrence. But ironically, the introduction of the notion of a causal model actually reduces the prospects in this regard, because it generates a need for a further argument to show how causal models are relative to the context of occurrence. Perhaps such an argument could be advanced (though at the cost of further complexity); but as it stands, Menzies' analysis is in the same boat in this regard as the contrastive strategy.[16]

4.7 Summary

Selection is slippery: it is not always easy to identify our intuitions about which event is *the* cause, even when we are entirely clear how a given effect came about (or at least, do not feel that *the* cause lies among events we do not know about or understand). Perhaps this is why the difference between cause and

[16]Another author expressing dissatisfaction with Lewis's gloss on difference-making is Carolina Sartorio. Her proposal states: "If C caused E, then, had C not occurred, the absence of C wouldn't have caused E" (Sartorio 2005, 75). This account counts the presence of oxygen as the cause of the match lighting: if the oxygen had been absent then the absence of oxygen would indeed not have caused the match to light. But Sartorio's discussion is not directed at the very general problem of selection have discussed, but at the much more specific problem of ruling out as causes events which merely switch the route by which a given effect happens (cf. Hall 2004a, Sartorio 2005).

condition has rarely been seen as a general, objective difference. Moreover we can be persuaded to talk with some flexibility about what *the* cause is. Perhaps this is why causal selection has seemed inessential to our concept of causation. There is no physical difference between cause and condition — no physical property possessed by all and only causes. Perhaps this, too, contributes to the view that there is no real difference, apart from our discriminatory treatment, between cause and condition, despite the obvious point that even a selective counterfactual theory of causation would not imply any physical difference between cause and condition.

Whatever the reason, causal selection has not received sustained and systematic attention in the way that other aspects of causation have. I identified three strategies which have been employed to understand selection: the special event strategy, the causal field strategy and the contrastive strategy. I suggested that they could usefully (though perhaps loosely) each be seen as refining the previous strategies. We discussed the contrastive strategy in detail, and I advocated two specific objections: that it fails fully to explain causal selection; and that it fails adequately to justify our moral and legal dependence on causal selection. Both the objections were presented as pressing the contrastive strategy to give an account of contrast choice.

The Reverse Counterfactual was introduced in the light of the notion of causes as difference-makers to their effects. I argued that the Reverse Counterfactual is true of causes, but false of mere conditions. I suggested that we could account for the flexibility of causal selection by the context-sensitivity of counterfactuals. Since counterfactuals depend on both context of occurrence and context of inquiry, I suggested we can explain why causal selection is relative to each of these. This yields an account which is more explanatory than the contrastive account, since it includes an account of the principles governing contrast choice. It also offers a better justification of the moral and legal weight we put upon selection, since it makes the principles governing selection important and quite general: they are just the principles governing our assessment of any counterfactual. The difference between cause and condition is not, therefore, one which is peculiar to moral and legal contexts, or peculiarly governed in those contexts. Another, more principled argument was advanced. Developing the idea that the Reverse Counterfactual underlies the vaguely-specified but frequently-invoked intuitive notion that causes make the difference to their effects, I sought to argue that the Reverse Counterfactual must in principle be true of causes, if causes really do make the difference

between the effect occurring and not.

The Reverse Counterfactual was compared to the contrastive strategy. It was argued that contrast still has an important role to play in explaining how the context of inquiry can sometimes determine that we select as causes events which, like background radiation, are absent only in rather distant possible worlds, and which are usually seen as mere conditions. The Reverse Counterfactual offers considerable theoretical unification, employing just one notion of context-sensitive difference-making, and explaining why contrastive explanation is such a powerful model of causal explanation. On the standard contrastive view, context determines the truth-value of the Lewisian counterfactuals, and then comes into play again in a separate way to determine our choice among those counterfactuals, all of which are true in the circumstances. On the Reverse Counterfactual account, context comes into play just once, to determine the truth-value of Lewis's counterfactuals and of the Reverse Counterfactual. Contrastive explanation is just a special way we have of specifying, manipulating and communicating context.

The obvious unanswered question of this chapter is how the Reverse Counterfactual might be incorporated or developed into a full theory of causation. This project will face several challenges. In particular, the Reverse Counterfactual is proposed as a necessary condition for causation. To achieve a sufficient condition, we would need new solutions to the problems associated with causal asymmetries — centrally, the difference between cause and effect, and the difference between effects of a common cause and cause-effect pairs. For Lewis's ban on backtrackers was his solution to those problems. These challenges will be taken up in Chapter 7; notice, meanwhile, that the selective feature claimed for the Reverse Counterfactual corresponds to a counterfactual asymmetry of sorts. The Reverse Counterfactual is true of many fewer events than the corresponding Lewisian counterfactual. Or, removing all presuppositions about temporal direction but retaining temporal order, we have many more counterfactuals going one way than the other, corresponding to the outnumbering of causes by conditions. This is a milder and subtler asymmetry than Lewis's. I hope that what we lose in an elegant account of causal asymmetries, we may gain in our understanding of the real nature of the asymmetry of counterfactual dependence. In Chapters 5 and 6 we shall learn more about the circumstances in which backtrackers are true, and find more reasons to accept the Reverse Counterfactual as a necessary condition for causation.

Chapter 5

Transitivity

5.0 Abstract

This chapter argues that the causes of the cause of an effect are not always causes of that effect: that is, causation is not transitive. I begin by distinguishing various kinds of nontransitivity. I consider four motivations for the common view that causation is transitive, and argue that none is compelling. Two sorts of counterexamples to transitivity are then distinguished: those due to *distance* (or failure of *proximity*) and those due to a special causal structure, *double prevention*. I argue that proximity failure is closely linked to causal selection. The Reverse Counterfactual tends to support the intuitive view that very distant events are not causes, even if they may be mere conditions. Then I argue, contrary to some recent literature, that cases of double-prevention fail to be cases of causation. The Reverse Counterfactual agrees. Efforts by McDermott, Paul and Hall to diagnose the difficulty with double prevention are considered and found wanting. However, I agree with these authors that double prevention does not constitute a real counterexample to causal transitivity; I suggest that only an unselective notion of causation could persuade us to think otherwise. Finally I consider how a valid substitute for counterfactual transitivity might help explain why causation sometimes appears to be transitive (unlike other nontransitive relations such as touching), and also why distance should be relevant to causal transitivity.

5.1 Kinds of Nontransitivity

A relation R is *transitive* just in case $(\forall x)(\forall y)(\forall z)((Rxy \& Ryz) \supset Rxz)$. A *nontransitive* relation (or a relation that is *not transitive*) is any which fails to

meet the foregoing condition, ie. any such that $(\exists x)(\exists y)(\exists z)((Rxy\&Ryz)\& \sim Ryz)$. Even if a relation R is nontransitive, it might fail to be *intransitive*: R is intransitive just in case $(\forall x)(\forall y)(\forall z)((Rxy\&Ryz) \supset\sim Rxz)$. If three cubes of the same size are placed so that the whole of one face of one cube is in contact with the whole of one face of a second cube, and the whole of one face of that second cube is in contact with the whole of one face of a third cube, then it is never the case that the whole of any face of the first cube is in contact with the whole of any face of the third (indeed, no faces will touch at all). The relation among cubes of the same size of *having faces aligned and in contact* is intransitive. Whereas the more commonplace relation of simply *touching* is nontransitive, but not intransitive: three cubes may be placed so that all three touch, or so that two touch a third without touching each other. I will argue that causation is nontransitive in the same way that simply touching is nontransitive: that is, causation is neither transitive nor intransitive.

5.2 Motivations for Transitivity

There is a tendency among philosophers in the post-Lewis tradition simply to insist that causation is transitive (cf. Hall 2004a,b, Ramachandran 2004), and it can be quite hard to get a grip on the motivations for the insistence and on the issues at stake in a debate on transitivity. In this section I shall identify and criticise four motivations for decreeing causation to be transitive. In each case I shall argue that the motivation is inadequate, especially given that the insistence on transitivity has led philosophers to neglect a more open-minded study of our intuitive judgements about transitivity and their governing principles.[1]

[1]Recent deniers of transitivity include Hitchcock [2001], who accepts counterexamples like those in 5.5.1 below. I shall not discuss Hitchcock's account in terms of structural equations, however. This approach seeks to develop a more sophisticated counterfactual approach which admits of causal "variables". This may be very sensible for certain practical and technical purposes (cf. Granger 1969), but it does not seem to help much with two fundamental issues concerning counterfactuals. First, the greater complexity threatens to obscure the difficulty of identifying the principles concerning what to hold fixed and what to change, in counterfactual suppositions (cf. Hall and Paul 2003). Second, Hitchcock's account assumes a ban on backtrackers, which I have already argued is unwarranted. For these reasons, despite the interest of Hitchcock's approach and related accounts such as Yablo's [2002], I feel that a satisfactory discussion would take us too far off course.

5.2.1 "Bedrock Datum" Intuition

The complexity of our causal judgements with regard to transitivity has not merely been ignored: it appears to have been denied. For example:

> That causation is, necessarily, a transitive relation on events seems to many a bedrock datum, one of the few indisputable a priori insights we have into the workings of the concept.
>
> (Hall 2004a, 181)

If this is an *a priori* insight then it is available only upon philosophical reflection, since ordinary thinkers show no evidence of it. Moreover, an *a priori* insight is quite different from an *a priori* argument: I might be convinced by an argument, but for someone like me who is blind to the insight, reports of it in others are not enlightening.

5.2.2 Accounting for Preemption

Paul says:

> [Regularity and counterfactual] accounts need transitivity to avoid important counterexamples...
>
> (Paul 2004, 206)

As far as counterfactual accounts go, she is presumably influenced by Lewis's theory of causation, which makes causation transitive in order handle preemption counterexamples. We will discuss Lewis's treatment of preemption in Chapter 6. Three points may be mentioned here, however. First, Lewis's theory and many others need causation to be transitive in order to handle preemption (cf. Paul 2004). But that doesn't mean that every other theory — not even every other *counterfactual* theory — needs causation to be transitive. Second, making causation transitive only provides a promising solution to preemption when combined with a denial of backtracking counterfactuals, as we shall see in 6.3.1. I have devoted two chapters to rejecting this denial; in the present context, they amount to independent reasons to reject Lewis's account of preemption, and thus to reject that account as a motivation for making causation transitive. Third, even if we grant Lewis's denial of backtrackers (as most do), none of Lewis's solutions to preemption, nor any other solution, is generally accepted as decisive. Preemption is widely considered to be the

most difficult and perhaps a fatal problem for the counterfactual analysis of causation. Proponents of particular solutions to preemption will argue that making causation transitive is a small price to pay for solving preemption. But since there is no agreement about how to handle preemption, this line does not constitute a general and independent reason to accept causal transitivity. The fact that transitivity is a feature of many different accounts of preemption provides little inductive support for the idea that it is necessary for a successful account of that sort. For these accounts are not very successful: an inductive argument is at least as viable from the failure of all such accounts to the rejection of transitivity.

5.2.3 Apollo

Lewis provides an explicit argument in favour of the transitivity of causation. Normally, we do not say that birth causes death:

> "Counterfactual analysis of causation? — Yeah, yeah, my birth is a cause of my death!" said the scoffer. His birth is indeed a cause of his death; but it's understandable that we seldom want to say so. The counterfactual dependence of his death on his birth is just too obvious to be worth mentioning.
>
> (Lewis 2004a, 101)

This is an instance of the view that causal selection is a matter of conversational, but not logical, implication. According to Lewis, to mention the birth as a cause of death would be inappropriate, but not strictly false.

In support, Lewis argues as follows. It has been foretold that your death will have catastrophic consequences for Apollo, who therefore orders an underling to prevent it. The underling chooses to do so by preventing your birth. However he fails. You are born, and die, with catastrophic consequences for Apollo. In that situation, it is natural for Apollo — and us — to consider your birth a cause of your death, and to curse his underling for failing to prevent your birth, for the reason that the failure led to your death. Now suppose that, as far as earthly affairs go, that world and ours are exactly similar. If we accept that your birth causes your death in that world, but deny it in the actual world, then you seem to accept that whether your birth causes your death depends on heavenly affairs. Lewis thinks that is obviously wrong:

> ...it would be entirely appropriate for Apollo to complain that your birth caused your death. And if it's appropriate to say, presumably it must be true. But now we may suppose that, so far as earthly affairs go, actuality and our unactualized comparison case are alike in every detail... So, if you agree with the scoffer that your birth didn't cause your death in actuality, you must think that idle heavenly difference can make a difference to what causes what here below! That is hard to believe.
>
> (Lewis 2004a, 101)

The argument is supposed to bring out that we can at least sometimes think of contexts where we override our initial judgements, and assert transitivity after all. This shows that apparent transitivity failure is a feature of context. Causation, in Lewis's opinion, is not a feature of context, but is intrinsic to the events involved. So we must say the same thing about the Apollo world and our godless actuality, since the same sequence of events leads from my birth to my death in both worlds. It is, Lewis assumes, more plausible to think that birth causes death in both worlds than to deny it in both worlds. And this assumption is surely right.

Even if the Apollo argument were persuasive, it seems unlikely that it is the reason why so many philosophers consider it at least acceptable to assume that causation is transitive; after all, that assumption appeared in Lewis's 1973 theory, and the Apollo argument was first published in 2000. Moreover this argument is not persuasive. It relies on two claims: (i) that my birth causes my death in the godly world; and (ii) that what goes for the godly world must go for the godless. Both claims may be disputed.

First, the only way to persuade me that my birth caused my death in either world is to specify my death extremely imprecisely. None of the features which my death actually has (whatever they may be) are caused by my birth, however. If we specify the death a bit more precisely — let us suppose it is death by thunderbolt — then little plausibility remains to the suggestion that my birth was its cause. The birth may have been necessary for the death, but the thunderbolt is what killed me. Nor will it help to specify my birth more precisely, assuming that my birth included no obvious omens of my demise: we will be little inclined to see any of the features of an ordinary birth such as mine, as causes of my dramatic death.

Second, we could deny that causation is intrinsic in the relevant way. In fact we need not even deny it: we need merely regard as suspect an argument which

appears to rely on the claim that causation is intrinsic, given that whether causation is intrinsic is a difficult and not entirely obvious question (cf. Menzies 2002). On my view it is far from clear that causation is intrinsic. For on my view causation is selective, and selection depends on context, which by definition fails to be intrinsic to the events in question. Lewis's argument brings out how his view and mine are at odds, and also how asserting transitivity and denying selection are strongly connected.

If this rather flippant attitude towards the instrincness of causation bothers you, note that Lewis's account also fails to make causation intrinsic, if my argument in 4.4.3 is correct. For all counterfactuals are relative to the context of occurrence; I count it as an advantage of Lewis's semantics that it brings this point out. Thus any counterfactual account of causation fails to make causation intrinsic to the events concerned, because whether one causes the other depends on the truth of a counterfactual, which in turn depends on the context in which the two events occur. And context fails to be intrinsic if anything does. Lewis denies he is relying on any strong thesis of intrinsicness; "all we need is that earthly causal relations supervene on the intrinsic and nomological character of all things earthly" (Lewis 2004a, 101). But we could still deny that: asserting that a relation is context-dependent amounts to denying that the relation under consideration supervenes on the intrinsic character of the relata. And for what reason would we rule that non-earthly things cannot be part of the context of causal relations — apart from the rhetorical pull which the word "earthly" might exert on a physicalistically inclined modern?

5.2.4 Opposition to Selection

The transitivity claimed for causation might be seen as a special case of the unselective nature of causation more generally. That is how Lewis sees it:

> We have the icy road, the bald tire, the drunk driver, the blind corner, the approaching car, and more. Together, these cause the crash... But these are by no means all the causes of the crash. ...each of these causes in turn has its causes; and those too are causes of the crash.
>
> (Lewis 1986a, 214)

If, like Lewis, you are committed to an entirely unselective notion of causation, then you might thereby be motivated to consider causation transitive.

On Lewis's view, a chain of counterfactual dependence suffices for causation.[2] Cases where we select the cause from mere conditions are counterexamples to this claim. So, too, are cases where we allow events in one part of the chain to count as causes, but not those in another part. As Hall puts it:

> ...transitivity helps to make for an egalitarian relation: Events causally remote from a given effect will typically not be *salient* — but will still be among its causes for all that.
>
> (Hall 2004b, 228)

Hall is clearly suggesting a connection between causal transitivity and an unselective — "egalitarian" — notion of causation.

Is there any difference, then, between endorsing transitivity and denying selection? There is, but for practical purposes, the difference does not matter. We could imagine a counterfactual view which did not identify causation with chains of counterfactual dependence, but with single steps — so $\sim C > \sim E$ was a necessary as well as a sufficient condition for c to cause e. Then, we would have an account which was unselective among mere conditions, but on which causation was non-transitive.[3] However, such an account is *prima facie* implausible, because then even mild cases of redundancy no longer count as cases of causation. The fact you were about to get up and put the lights just when I did, means that my pressing the switch did not cause the lights to come on: it is false that, if I hadn't pressed the switch, the lights wouldn't have come on. So although opposition to selection does not entail transitivity, for practical purposes it might as well, because denying transitivity while endorsing an unselective notion of causation makes for an extremely unappealing view of causation which, to my knowledge, nobody holds.

This explains why denying selectivity and endorsing transitivity are so closely linked. It hardly needs to be pointed out that this is not, by my lights,

[2]Strictly, a chain of causal dependence is required; but I do not use this terminology, because it is biased against my view. As a reminder, e counterfactually depends on c when $C > E$ and $\sim C > \sim E$, and Lewis stipulates that e causally depends on c when in addition c and e both occur. So counterfactual dependence plus C and E entails Lewisian causal dependence, and Lewisian causal dependence entails counterfactual dependence on its own. The bias arises because causal dependence strongly suggests causation; whereas I deny that events which are causally dependent in Lewis's sense need instantiate causation. On the other hand, to define my own version of causal dependence would invite confusion. So I eschew the terminology of causal dependence.

[3]But not intransitive: for sometimes, c is a condition of d and d of e, but c is also a condition of e. For instance, my birth is a condition of my drinking a cup of coffee this morning, which is a condition of my perking up; and my birth is also a condition of my perking up. So a view on which $\sim C > \sim E$ is necessary for causation rules out intransitivity.

a good reason for endorsing transitivity; for in Chapter 4 I argued against the view that causation is unselective among conditions.

5.3 Two Sorts of Transitivity Failure

There are at least two sorts of case where causation might appear not to be transitive. One sort is relatively straightforward to describe roughly, if not to characterise precisely: some event c seems to be too distant from some causal descendant e for us to call it the cause. If there really are cases like this, then they constitute direct counterexamples to the alleged transitivity of causation. This is the sort of case which centrally concerns Hart and Honore, in legal contexts; they want to uncover principles by which we draw the line between events that are sufficiently proximate to count as a cause, and event that are too remote (see especially Hart and Honore 1985, Chapter V). Let us call this kind of transitivity failure *proximity failure*, and counterexamples due to proximity failure *distance counterexamples*.[4]

The other sort of case exhibits a more complex causal structure: specifically, *double prevention*. This is the sort of case the Lewis-inspired literature focuses on. c causes d and d causes e, but c does not cause e; and moreover, the failure does not seem to be because c is distant from e. The special feature of double prevention cases is that c on its own would prevent e from happening; but then d prevents c from preventing e. And the occurrence of d counterfactually depends on the occurence of c. Paradigm cases are failsafes which actually fire: the power supply to the hospital is disrupted, so the emergency generator kicks in, causing the power to stay on. Although there is a clear chain of counterfactual dependence, we do not normally say that the continued power supply at the hospital is caused by the failure of the national grid.[5]

The two sorts of cases are seen as distinct, and the latter lends itself to the counterexample culture, because once you know the double-prevention for-

[4]Pending a better suggestion, and without anything depending on it, the notion of distance may as well be spatiotemporal. (This diverges somewhat from legal usage, where proximity can simply be a technical term for limiting liability within a set of causes — the American version of the English distinction between "factual" and "legal" causation.) It is often said that there should be no causal action at a distance, and perhaps this intuitive principle is part of what motivates distance counterexamples. Note, however, that even if this is part of the motivation, distance counterexamples to transitivity need not strictly violate the principle, if it is formulated to allow causation between events connected by causal chains. We are interested in events which *are* connected by causal chains, but which — due to the length of the chains — are not naturally regarded as standing in a direct causal relation.

[5]The point about failsafes and the generator example are due to Peter Lipton.

mula it is fairly easy to generate examples. The most-discussed problems for Lewis's theory of causation take the form of counterexamples: cases of preemption, which challenge Lewis's account as a necessary condition for causation. Transitivity is a central focus of the more recent attention to the sufficiency of Lewis's account, and given the counterexample culture that has developed in the discussion of preemption, perhaps it is natural that discussion of transitivity should also be cast in terms of counterexamples. But this has had an unfortunate consequence: little attention has been paid to the principles underlying our intuitions, and the way these generate counterexamples.[6] What attention has been paid, has been secondary, prompted by defence of transitivity. Given the transitivity claimed for causation, it is perhaps surprising that little positive interest has been taken in why we fail to possess strong intuitions that causation is transitive (and perhaps the lack of intuitions has been forgotten — see 5.2.1 above). The explanation, I think, is that the defence of transitivity has been seen as a defence against counterexamples — central to an activity which Hall, following Maudlin, calls "trench warfare" (Hall 2004b, 227). Whereas we might seek to articulate the principles governing our intuitions about transitivity, as well as merely defending theoretical accounts from those intuitions.

The general goal of this chapter is to argue that causation is not transitive, and we shall begin with the first sort of failure of transitivity distinguished above — distance failure.

5.4 Proximity Failure

> It is to be noted that, despite what is commonly said by philosophers, causal relationships are not always 'transitive': a cause of a cause is not always treated as the cause of the 'effect'... Thus the cause of a fire may be lightning, but it would be rare to cite the cause of the lightning (the state of electric discharges in the atmosphere) as the cause of the fire; similarly the cause of the motor accident may be the icy condition of the road, but it would be odd to cite the cold as the cause of the accident.
>
> (Hart and Honore 1985, 43)

To the extent that Hart and Honore's example of the ice and the road is plausible, the Reverse Counterfactual has the resources to match our intuitions.

[6] Perhaps this is changing: a recent exception is Björnsson 2007.

Applying the Inference Test: if the car had not crashed, then we might have inferred there was no ice; but it is less obvious whether we would infer anything about the temperature. This indicates that our confidence in —

> If there had been no crash, there would have been no ice,

is (or should be) greater than in —

> If there had been no crash, it would not have been cold.

Admittedly this is rather uncertain, but that reflects the limited details supplied, and also the fact that our insistence on calling the ice, and our resistance to calling the cold, the cause of the crash is perhaps not so strong as Hart and Honore seem to think. (Note that it is the same example Lewis appeals to in the passage quoted on page 111, when pumping our intuitions in the opposite direction.) Nevertheless, to the extent that — and in the circumstances that — we agree with Hart and Honore's view, the Reverse Counterfactual agrees too. In support of the claim, I employ the Inference Test: I suggest that our inclination to call the cold (or the ice) the cause of the crash, will vary with our inclination to assent to the inference from supposed absence of crash to the supposed absence of cold (or ice).

Turning to a more common and perhaps more convincing example, my birth does not cause my death, in the ordinary sense: it does not kill me. Yet on Lewis's analysis, my birth is a cause of my death, since if I had not been born I would not have died. The Reverse Counterfactual, on the other hand, does not count my birth as causing my death. Suppose I die in a car accident, caused by a drunk driver, who could also be me. Legal and moral consequences ensue for the drunken driver, even if it is me, since others may have a claim for compensation against my estate. Nobody, however, has a claim against my mother, simply for having brought me into the world. If the event of my death had not occurred, then there is no obvious reason to suppose the event of my birth would not have occurred (nor even that it would have occurred differently); but presumably the car would have been travelling a bit slower, or I would have reacted more quickly or deftly, or I would not have been so drunk; or perhaps I would have taken the train, or some such. The Reverse Counterfactual will identify the event which makes the actual case differ from the counterfactual case where I do not die, as the cause of my death, selecting them from other events in the chain that leads to my demise. There are many

other deaths I could have died; supposing me unborn is by no means the only or the closest alternative to the death I actually die.

To make the mode of argument explicit, the Inference Test confirms that if I hadn't died, the car wouldn't have crashed as it did. If we suppose that I arrived at my destination alive, we would naturally infer (assume, even) that no crash had occurred.[7] On the other hand, there is no support for the thought that if I hadn't died that death, I wouldn't have been born. Suppose I had survived; that would clearly not be a natural reason to suppose I had never been born. And although supposing me out of existence entirely is perhaps a minor change in the greater scheme of things, it is still a larger change than allowing me to live a little longer, shorter or differently from how I actually do.

This is perhaps a useful point to note how weakly the Inference Test needs to be applied, in order to support the argument. I do not depend on any very strong claim either about what we would or would not infer, under certain counterfactual suppositions; all I really need is a clear *asymmetry* between these things. In the case where we compare my birth to my drunken driving as the cause of my death, that asymmetry survives minor quibbles about exactly what we infer from the absence of my death. The Reverse Counterfactual therefore appears to offer a way of ruling out my birth as a cause of my death, at least in the common sorts of cases where we want to do so.

There remains some context-relativity, as we have already noted. The medics might say that the cause of my death in the car crash was brain-damage. But the Reverse Counterfactual account accommodates this sort of context-dependence, by allowing that the vagueness of counterfactuals is resolved differently in different contexts of inquiry. As I argued in the last chapter, resolving this vagueness is quite different from choosing between true counterfactuals whose vagueness has already been resolved — which is the status of selection on the Lewisian counterfactual analysis. Resolving counterfactual vagueness is necessary for any counterfactual account of causation: on my account, therefore, selection is integral, and no less objective than causation itself will be on any counterfactual account.

The Reverse Counterfactual appears to offer a principled account of the nontransitivity of causation. When c causes d and d causes e, c further causes e only if $\sim E > \sim C$. The principled argument of 4.5.4 applies here too. For

[7] As with the match-strike example of the previous chapter, we might not infer quite that much — we might allow that a small crash could have taken place. But this just demonstrates an obvious point: that some of the crash's properties are causally relevant.

some c to cause some e, c must be a difference between the actual case and the counterfactual scenario where e does not occur. Very remote events satisfy this condition less commonly, perhaps partly because — as Lewis's own view of counterfactuals implies — the counterfactual scenarios we conjure up to assess our counterfactuals tend not to alter history more than they need to. Of course, Lewis's view about how much history needs to be altered differs from my own. Nevertheless, if the Reverse Counterfactual can be defended in the way I am defending it, by appeal to the Inference Test, then it seems that endorsing backtrackers need not mean abandoning counterfactual asymmetry altogether.[8] Rather, we have uncovered another feature of the limited asymmetry of counterfactual dependence which we are exploring as a sort of side-effect of our discussion of causation.

5.5 Double-Prevention

5.5.1 Introducing Counterexamples

One advantage of the obsession of the modern literature with counterexamples has been the rise to prominence of the relation between double-prevention and transitivity. This is a special instance of the more generally important relation between transitivity and redundancy. I shall characterise double prevention more precisely below, but the general problem is supposed to be this: there is a chain of dependence from c to e, yet intuitively c fails to cause e. In this section, we shall seek to establish two things. First (5.5.3) I shall argue that in cases of double prevention, the Reverse Counterfactual agrees with our intuitions that c does indeed fail to cause e (5.5.2). This relieves us of the duty some writers have felt to argue against our intuitions. Second (5.5.4) we shall consider three other efforts to understand double prevention cases and their relation to transitivity. Each fails, but I suggest that what is good about them is captured by the Reverse Counterfactual account.

This section thus completes the main work of the Chapter, since it completes the defence of nontransitivity against various accounts which have been devised in defence of the claim that causation is transitive. But it leaves the nature of double prevention undiagnosed: in particular, we might ask whether cases of double-prevention really are cases of transitivity failure. More generally, we might ask why causation was thought to be transitive, and why that

[8] And thereby endorsing the implausible view that counterfactuals imply entire counterfactual histories, as well as, or — worse — instead of, counterfactual futures.

suggestion has some intuitive pull. I address these questions together in 5.6.1, and offer a diagnosis of proximity failure in 5.6.2.

Now let us consider some examples of double-prevention which occur in the literature (and which I have slightly adapted by inserting the author as the protagonist where a name is not supplied). These are the examples considered by the three authors whose discussions of transitivity we shall focus on.

Dog-Bite. McDermott has occasion to detonate a bomb (McDermott 1995, 531). The day before he does so, a dog bites off his right forefinger. So he detonates the bomb by pushing the button with his left forefinger. If the dog-bite had not occurred, then the left-handed button-push would not have occurred, and if that left-handed button-push had not occurred, the bomb would not have detonated. If causation is transitive, and if the truth of each of these counterfactuals suffices for causation, it follows that the dog-bite caused the explosion. Yet intuitively, the dog-bite does not cause the explosion.

Shock C. McDermott also invents a sadistic game called Shock C, played by an unruly pair, Able and Baker (thinly disguised in McDermott's account as A and B). A poor subject, called C, is wired up to an electric circuit with two switches. Able controls one, Baker the other. When the switches are both set the same way, C receives a shock; when they are set differently, C does not receive a shock. Able's aim is to prevent C being shocked, and Baker's is to shock C. Able sees Baker's switch set left, so throws his right. Baker, seeing Able throw his switch right, follows suit. The switches are aligned; C receives a shock. McDermott says that "Common sense tells us that A's move was a cause of B's move, that B's move was a cause of C's shock, but that A's move was not a cause of C's shock" (McDermott 1995, 532).

Chest Massage. McDermott gives a chest-massage to a heart attack victim, saving his life. The victim lives to travel to New York and die another day. Sharing the preoccupation which seems to infect his profession, McDermott specifies that the victim dies violently[9] (McDermott 1995,

[9]This obsession is sometimes remarked upon (eg. Hall 2004a, 183) in passing. But recent psychological research suggests that the choice of examples is of greater significance: "Counterfactual thoughts focus on specific antecedents that could inhibit a bad outcome, whereas causal explanations focus on both general and specific factors" (McEleney and Byrne 2006, 247). Investigating the differences between causal and counterfactual reasoning would be an intriguing avenue of explanation for some philosophical difficulties that have beset counterfactual analyses of causation.

532). The massage is a cause of the journey to New York, which is a cause of the death, but the massage is not a cause of the death, says McDermott.

Skiing Accident. Right-handed Suzy goes skiing, breaks her right wrist, writes a philosophy paper with her left hand, and submits it to a journal where it is accepted for publication. The skiing accident is a cause of her writing the paper with her left hand, and the writing of the paper with the left hand is a cause of the publication, so by transitivity, the skiing accident is a cause of the publication. Yet our intuitions disagree (Paul 2004, 205).

Dog Yelp. Suzy prepares to throw a water-ballon at a dog. Billy runs to stop her, but trips over a tree-root. The balloon hits the dog, which yelps. Billy's run caused his trip, and the trip caused the dog's yelp (assuming if suffices for causation that, if Billy hadn't tripped, the dog wouldn't have yelped). Thus, if causation is transitive, Billy's run caused the dog to yelp. Yet we do not want to say that the run caused the yelp (Hall 2004a, 183–4).

Discovered Bomb. An assassin plants a bomb under Hall's desk, Hall finds it and removes it, and thus survives. Planting the bomb caused Hall to find it, finding it caused him to survive, but planting the bomb did not cause Hall to survive (Hall 2004a, 183 — crediting Field).

Kvart's Finger. Kvart's finger is severed in an accident, and sewn back on so well that a year later, it is as good as it ever was. The accident caused the surgery, the surgery caused the finger's health a year later, but the accident did not cause the finger to be healthy a year later (Hall 2004a, 183 — crediting Kvart [1991]).

5.5.2 The Structure of the Counterexamples

Each of these cases has the following structure:

- c occurs;
- d occurs, and counterfactually depends on c;
- e occurs, and counterfactually depends on d;
- intuitively, c does not cause e.

In my opinion, they also possess two features which cannot be characterised as easily:

- they are all cases where c prevents some event d', which would cause e; but where
- d causes e instead of d', thus preventing c from preventing e.

For example, take the Dog Bite. McDermott's finger getting bitten off prevents him from pushing the button with that finger, which would detonate the bomb. But his left-fingered push does the job the right-fingered push would have done. The left-fingered push thus prevents the bite from preventing the detonation, which why this structure is called *double-prevention*.

This diagnosis is close to one given by Hall:

> An event c occurs, beginning (or combining with other events to begin) some process that threatens to prevent some later event e from occurring (call this process "Threat"). But, as a sort of side-effect, c also causes some event d that *counteracts* the threat (call this event "Savior"). So c is a cause of Savior, and Savior — by virtue of counteracting Threat — is a cause of e. But — or so it seems to many philosophers — c is not thereby a cause of e, and so Transitivity fails.
>
> (Hall 2004a, 184)

Notice that I have weakened Hall's picture, since I have not required that c (which Hall calls "Threat") causes d (which Hall calls "Savior") and that d causes e. For double prevention cases to be a counterexample to the transitivity of causation, we need Hall's stronger version. But for cases of double prevention to be counterexamples to the claim that counterfactual dependence suffices for causation, the weaker picture is all we need. In this section, we are considering the latter question — that is, we are considering how transitivity is a problem for a Lewisian theory of causation. In that tradition, it is widely assumed that counterfactual dependence suffices for causation, so the distinction is often blurred; but we must make the distinction because we have questioned that assumption. In the next section we shall consider the more general question, whether double-prevention cases are counterexamples to causal transitivity.

5.5.3 The Reverse Counterfactual Gets It Right

I endorse each of these counterexamples, in the sense that in each, I agree that the first event does not cause the final event. And the Reverse Counterfactual reflects this fact. In Dog Bite, if the bomb had not been detonated, then McDermott would not have pushed the button — with any finger. Suppose you knew of McDermott's plan, and noticed that no explosion occurred at the appointed hour. You would infer that he had not pressed the button. — Unless, of course, he is an avid but incompetent terrorist, whose bombs routinely fail to explode; but in that case, the button push and the well-made bomb jointly cause the explosion, and if the bomb does not explode, you cannot decisively infer the absence of either rather than the other. Such caveats should, by now, be growing familiar. Contrast the dog-bite: you would guess nothing about a dog-bite from the failure of a bomb to explode; certainly, you would not infer that a dog had *not* bitten his finger off on that basis. It is false, then, that if the bomb hadn't exploded, the dog would not have bitten off McDermott's finger. The dog-bite fails to be a cause of the explosion, according to the Reverse Counterfactual, as it does according to our intuitions.

The case of Shock C is less clear. I have some doubts about the example: if the current is switched on *before* Able moves, then C will receive a shock, and the game will be over before Baker moves. So it must be switched on either after Able moves or at exactly that moment. Then, I think, it is intuitively clear that Baker's move causes the shock, and that Able's does not. Some time elapsed between Able's move and Baker's, during which time, the power was on, and C could have been receiving a shock. Then Baker moves, and C is shocked. We could have intervened to prevent this, by overpowering Baker or switching off the current. It is hard to see Able as causing the shock. Those are our intuitions, and the Reverse Counterfactual agrees: in the absence of a shock, we would infer that Baker had not moved, given the details we have just filled in concerning the timing of the current being switched on relative to Able's move. Of course it once again depends how much we are supposed to know about the circumstances — crucially, whether we hold Able's switch fixed or not. But this reflects our causal intuitions: after all, it is only because we are told about Able's move, and the fact it breaks the circuit, that we say that Baker's move caused the shock; if the case were less precisely specified, our causal intuitions would not be as clear, if clear at all.

The case of the Chest Massage failing to cause a subsequent trip to New York, and the ensuing violent death, is also handled by the Reverse Counterfac-

tual. A different, peaceful death from the actual violent end would not license an inference that no chest-massage had taken place. (Here, we are revisiting our discussion of proximity failure: the chest massage is just too distant to count as a cause of the subsequent death, even if it is a mere condition.) Therefore the Inference Test undermines any claim that the chest-massage meets the necessary condition imposed by the Reverse Counterfactual for causing death. So our account rules out the chest-massage as the cause of death.

Paul's Skiing Accident is very similar to McDermott's Dog Bite, in that the first event in the chain bears on aspects of the second event which are causally irrelevant to the final event. Paul takes her cue from this point, as we shall see. But here we need only note that the Reverse Counterfactual treats it successfully, as it treated Dog Bite. If the paper had not been accepted, we should not infer that no skiing accident had taken place. Whereas we could infer from a rejection that the paper was not of sufficient quality, though perhaps not with overriding conviction. But our lack of conviction reflects a suspicion that there are non-meritocratic elements in the editorial policy of academic journals, which might include eminence of author, chance, pressure on publication, poor refereeing, and so on. These are all factors which diminish both our confidence in the outcome of the Inference Test, and our confidence that the Reverse Counterfactual is true of the writing of a high-quality paper with respect to the subsequent publication. But that is no objection: for the presence of anti-meritocratic factors equally diminishes our confidence in the notion that writing a good paper causes publication on a given occasion, without knowing a lot more about the process by which this particular publication was achieved. (Indeed, a vehement case is sometimes made that writing a good paper is no more than a mere condition, and perhaps not even that.) Thus our causal convictions once again vary in tandem with the conviction which the Inference Test recommends we should have in the Reverse Counterfactual.

In the Dog Yelp case, the problem is supposed to be that intuitively, we deny that Billy's run caused the dog's yelp, despite agreeing that if he hadn't run he wouldn't have tripped, and if he hadn't tripped the dog wouldn't have yelped. But in this case, we have an extra step in the chain: we also have Suzy's throw, upon which all the other events counterfactually depend. Moreover, intuitively we are inclined to count Suzy's throw as causing the yelp. So we need to check both that the Reverse Counterfactual correctly discounts the run as the cause of the yelp, and that it correctly counts the throw as at least among the causes. So: applying the Inference Test to the situation where the

dog doesn't yelp, we might infer that Billy didn't trip. But then we would infer that he *had* still run, but simply avoided the tree root. And we would infer that Suzy didn't throw — Billy succesfully prevented her. So both the trip and the throw count as causes of the yelp, but the run does not. The trip and the throw are both differences between the actual case and the most obvious counterfactual scenario where the yelp does not occur; but the run is present in both actual and counterfactual cases.

In another context of inquiry (perhaps specified by explicitly drawing a contrast) we might infer that Suzy didn't ever make as if to throw in the first place. In that case, Billy would neither have run nor tripped. Does the Reverse Counterfactual thereby wrongly count those events as causes of the yelp, in that context? There is indeed a problem here, but a different problem: this is a case of *spurious causation*. For the difficulty is that Suzy's throw is a condition for Billy's activities as well as for the dog's yelp. This is a different issue from the nontransitivity of causation, and will be addressed in Chapter 7.

Now remember Discovered Bomb, which an assassin plants and which Hall subsequently discovers under his desk. The Reverse Counterfactual says: if Hall hadn't survived, the bomb wouldn't have been discovered. Why? — Because, on hearing that Hall had perished in a terrorist attack, we would infer that he probably did not discover the bomb before it exploded. (If he did, why didn't he do something about it?) On the other hand, we would not infer that the assassin had not planted a bomb: quite the opposite. So the assassin's action fails to count as a cause of the survival, whereas Hall's finding the bomb does count.

Finally consider Kvart's finger, injured in a factory accident and, following excellent surgery, healthy a year later. The Reverse Counterfactual pronounces the surgery a cause of the health, surely: if the finger had not been healthy a year later, that would be because the surgery hadn't been so good. But it would not be because the accident never occurred in the first place. The application of the Inference Test is becoming routine: we could infer bad surgery, but certainly not an accident-free past, from a less healthy finger a year later.

We could extend this defence to other examples, but I hope the pattern is now clear. Let us run it by the example which I said was paradigmatic of the double-prevention structure, and a good deal simpler and more obvious than some of the previous examples. There is a power-cut, causing the hospital generator to kick in, leaving the power supply at the hospital uninterrupted

(or, in any case, on). If the national grid hadn't failed, the generator wouldn't have kicked in; and if the generator hadn't kicked in, there would have been no power supply subsequently; but the grid failure does not cause the subsequent power supply. The Reverse Counterfactual says so too. If the power at the hospital had been down, the generator would not have been on, presumably; but it is absurd, surely, to try to argue that if the power at the hospital had been down, there would have been no power-cut. For then, why would the power supply at the hospital have been down? So the generator, but not the power-cut, qualifies as causing the power supply.

5.5.4 Other Accounts Get It Wrong

Let us consider three instructively erroneous efforts to defend causal transitivity from the foregoing alleged counterexamples. McDermott suggests that a counterfactual sufficiency account of causation might help — that is, an account on which causes are counterfactually sufficient for their effects. He is right: it does; although the account he proposes suffers some difficulties. The Reverse Counterfactual offers a better account of this sort. Paul suggests that *aspects*, rather than events, be taken as the causal relata (where aspects are property instances). This move highlights an important feature which some alleged counterexamples to transitivity have. But they do not all possess this feature, and besides it is unclear that how to cash out her suggestion. The Reverse Counterfactual explains why aspects are sometimes relevant to questions of transitivity. Finally, Hall suggests that in some cases, we should reject the claim that counterfactual dependence suffices for causation. Quite so: we should; for in many cases, effects counterfactually depend upon events which are mere conditions for them, and not causes. But we should reject dependence in a wider range of cases than Hall suggests. — Thus each of these three accounts includes components which I want to endorse.

McDermott's Sufficiency Thesis

McDermott's proposal is that we identify minimal sufficient sets of events as direct causes, and chains of direct causes as indirect causes. A sufficient condition for an effect e is a set of events such that, if any events outside the set had failed to occur, e would still have occurred. A minimal sufficient condition is just a sufficient condition which contains no sufficient conditions as proper subsets (McDermott 1995, 533). A causal process is defined as a

chain of minimal sufficiency leading from c to e. For c to cause e, there must be a causal process running from c to e, and further, c must be a minimal condition for the *nominal essence* of e. The nominal essence of an event is the way we refer to it; the *real essence* is "a full intrinsic description of the event, including a precise specification of the time of occurrence" (McDermott 1995, 540).

Consider the Chest Massage case. McDermott accepts that "there was a causal process P leading from the massage to his death" (McDermott 1995, 543); however, he claims that the chest massage is not a minimal sufficient condition for the nominal essence of the death in New York, at least under certain ordinary descriptions. When we refer to the death, we specify the event loosely enough to allow that it could have come about without the chest-massage, because we speak loosely enough to allow that it could have come about without the earlier heart-attack as well — even though the chest massage cannot be eliminated from a minimal condition for the real essence of the death.

It seems to me that this distinction between real and nominal essences undercuts some of the motivation for McDermott's account, which was to handle causal redundancy. Simple cases of redundancy seem to be counterexamples: if there is a back-up assassin, then the primary assassin's firing at the president is not part of a minimal sufficient condition for the nominal essence of the president's death (whose nominal essence is just that — the president's dying). It may stand in that relation to the *real* essence, but then the chest massage is part of a minimal sufficient condition for the *real* essence of the death in New York (remove it, and there is no death in New York: the victim dies of a heart attack years earlier). I am not clear how McDermott might respond to objection.

But despite these doubts, the account serves to highlight an important point. Double prevention involves redundancy, in the sense that the two preventers taken together are redundant, cancelling each other out as it were. We have already noted that redundancy is *prima facie* a problem for accounts which make causes counterfactually necessary for their effects, and that its bearing on sufficiency accounts is less obviously problematic. We will develop this point in Chapter 6. The Reverse Counterfactual clearly does not make causes counterfactually necessary for their effects; it makes them counterfactually sufficient.

Paul's Aspect Thesis (and Anscombe on Extensionality)

McDermott's account is inferior to the Reverse Counterfactual in two other important ways. It is more complex, and it has nothing to say about proximity failure. Paul has an account which is better in these respects. The central suggestion of Paul's solution is to make event *aspects*, rather than events, the causal relata. An aspect is a property instance (Paul 2004, 213). Paul then seeks to combine counterfactual and regularity theories to yield an analysis of causation, taking aspects as the relata, which handles preemption and preserves transitivity.

Paul's central idea is simple and intuitive. In her Skiing Accident, the arm-break affects an aspect of the writing of the paper which is causally irrelevant to whether the paper is published, even though it was that very writing which caused the paper's publication, and that very writing was left-handed. Moreover, Paul's aspect thesis might shed some light on proximity failure. The longer I live, the fewer aspects of my death plausibly depend on aspects of my birth, even if my death depends on my birth. As far as I am aware, Paul does not discuss this possible application, but at least it seems that she might.

Paul's solution suffers from two severe difficulties, however. First, it lacks generality. Hall notes this, pointing out that the solution will not generalise even to fairly close counterexamples:

> Suppose that after the dog-bite, the man does not push the button himself but orders an underling to do so. The relevant intuitions do not change: the dog-bite causes the order, and the order causes the explosion, but the dog-bite does not cause the explosion. The only way I can see to apply Paul's observations is by way of a rather strained insistence that there is one event — call it a "making the button depressed" — which the dog-bite causes to have the aspect "being an order," and which otherwise would have had the aspect "being a button-pushing."
>
> (Hall 2004a, 187)

Hall is being kind. This solution is not merely "unattractive". It also undercuts much of Paul's motivation: for such a "strained insistence" that there is one event which can possess the different aspects singled out by Hall, carries heavy commitments concerning the nature of events, which was something Paul set out to avoid.

Secondly, Paul's substantive suggestions about how to construct a theory of causation based on aspects are not promising. She countenances two ideas: influence and lawful entailment. Neither seems well-suited to hold between aspects. Influence between *events* might be roughly characterised as counterfactual dependence between the aspects of the events in question: dependence of how on how and when on when, as well as whether on whether (Lewis 2004a, 91). What, then, is influence between *aspects*? The "how" and "when" of an event are aspects of that event. If aspects influence each other, then do they in turn have aspects? Aspects are property instances, and I suppose a property instance could itself instantiate various second-order properties, but Paul doesn't give any details, so it does not appear likely that this is what she has in mind. But if aspects don't themselves have aspects, then it is hard to understand how they can stand in "how-how" and "when-when" dependencies. For, as property instances, they have their hows and whens essentially. Perhaps I misunderstand Paul here: but she does not provide much guidance.[10]

It is also unclear how aspects are supposed to lawfully entail other aspects. On its own, a single aspect will not lawfully entail much; fill in all the nomologically independent details, and under determinism it will entail everything that's left. For instance, if it is genuinely the last remaining independent detail, then by determinism, Paul's body temperature will after all determine the colour of the paper in front of her (Paul 2004, 219). For illustration, imagine that we set her body temperature at a value of several million degrees: then the paper will ignite, and so presumably change colour. On the other hand, if we consider the lawful entailments of an aspect relative to some proper subset of all the logically independent details, then we face a problem analogous to the one we had hoped to escape. For there will be some sets such that the left-handedness of Suzy's writing will lawfully entail the publication of the paper. The problem of working out what to specify and what to leave out is directly analogous to the problem of characterising events more or less precisely, by appeal to some properties rather than others.

So Paul's point that transitivity sometimes fails when causally irrelevant aspects are involved does not merit shifting our discussion of event-causation to a discussion of aspect causation. For it is *only* sometimes. Moreover, aspect causation appears not to have all the resources available to events-based analysis (though of course, for all I have said, better aspect-analyses might be forthcoming). Nevertheless, transitivity does *sometimes* seem to fail

[10]We will discuss influence in more detail in 6.4.3.

for the reason Paul highlights, and the selectivity of the Reverse Counterfactual captures this point. We might infer the absence of a writing from the lack of publication, but since it is irrelevant which hand wrote the paper, we will not infer anything about that; so we will have no reason to infer anything about events which might determine which hand the paper was written with, such as a skiing accident.

In fact, Paul's real point has nothing to do with transitivity. It may be brought out with single causal steps. Compare:

> Suzy wrote the paper with her left hand because she broke her right wrist in a skiing accident;

and:

> Suzy wrote the paper because she broke her right wrist in a skiing accident.

The first of these explanations is true, and the second is false. But these are causal explanations; what makes the first one good but the second bad is that the first, but not the second, identifies a cause of the publication.

In 5.6.1 below, I shall argue that all cases of double-prevention share this feature. But I have already agreed with Hall that not all cases of double-prevention succumb to Paul's aspect analysis. The explanation, I think, is that Paul has noticed *one way* in which a *sine qua non* for an effect can clearly fail to be its cause. I suggest that this is a special case of a more general phenomenon, namely, the selectivity of causation.

I said that Paul's real point has nothing to do with transitivity. What is her real point then? — I suggest that it is another way of putting the point Anscombe was making when she said that causal contexts are not extensional. Paul's example has a similar structure to one of Anscombe's:

> There is an international crisis because the man with the biggest nose in France made a speech.
>
> (Anscombe 1969, 155)

That is silly, of course; the crisis was because the President of France made a speech. But the example has the same effect as Paul's: by picking out an event by a causally irrelevant aspect, we create causal statements which are intuitively false, or at best very strange. The chief difference between the two

examples is that in Paul's case, the left-handed writing entails the writing: the causal event is specified by entailment. Whereas in Anscombe's case, the relation is not one of entailment, but of contingent fact: de Gaulle happened to have the biggest nose in France at the time. I do not think this feature is essential to the point underlying both examples.

Anscombe goes on:

> ...one... says, ... "But not *because* his is the biggest nose." — Now of course those who believe causal statements to be extensional will give an account of the "greater explanatory force" of the second member of each trio. But the question here is not whether one can defend a thesis through thick and thin... but really whether there was originally any good reason for this thesis at all. Here I am in a bit of difficulty. For I have no sure insight into the source of the conviction that causal statements are extensional.
>
> (Anscombe 1969, 155)

Maybe we can help. Paul's example only works in a climate where causal selection has been rejected. We might, as Anscombe says, acknowledge that saying "she wrote the paper with her left hand" is less explanatory or perhaps even misleading; but if we accept the unselective orthodoxy, then as far as the facts go, the writing of the paper with the left hand is a cause of the publication. Perhaps, then, the motivation for the conviction that causal contexts are extensional is a commitment to an unselective notion of causation. An unselective notion will force us to accept the speech of the big-nosed man as a cause of the crisis: for that speech was a *sine qua non* of the crisis. If I am right that there is a connection here, this is another reason to doubt the orthodox view of selection; for Anscombe's discussion of the intensionality of causation is well-known, and her argument that causal contexts fail to be extensional is rather convincing. Yet it is hard to see how we could reject the extensionality of causal contexts, with Anscombe, while keeping the doctrine that causation is unselective, with the Mill-Lewis tradition. I will generalise this point in 5.6.1, and argue that in every double-prevention case, only an unselective conception of causation would tempt us to see the first preventer as causing the second.

Hall on Rejecting Dependence

Hall argues that "double-prevention is not causation", and as I keep promising, I will endorse this view in the 5.6.1. However, my reason is different from Hall's. His goal is to argue that transitivity conflicts with the *dependence thesis* — the sufficiency claim that if e counterfactually depends on c, then c is a cause of e (Hall 2004a, 181). He thinks that double-prevention cases show this conflict, because we can resolve them by dropping either transitivity or the dependence thesis. I endorse the suggestion that we reject dependence whole-heartedly, but I think we also need to reject transitivity.

However, Hall does not think that all the cases we considered motivate the rejection of the dependence thesis. He thinks that Dog Yelp does, but that Kvart's Finger, Dog Bite and Discovered Bomb do not. In each of these cases, Hall thinks we should bite the bullet and accept that causation is transitive, despite intuitions to the contrary. The distinction between the two sorts of case is that, in one, there is a causal process connecting the first event to the last, but in the other, there is not. An example of the latter sort is the Dog Yelp case:

> This last example is an Easy case — Easy because it is so obvious how to respond to it in a way that safeguards *Transitivity*. After all, the only sense in which Billy's trip "causes" the dog's yelp is that it prevents something — Billy continuing to run towards Suzy, reaching her in time to stop her from throwing the balloon, and so on — which, had it happened, would have prevented the yelp. But no *causal process* connects the trip to the yelp...
>
> (Hall 2004a, 184)

And this easy answer amounts to rejecting the claim that counterfactual dependence suffices for causation. In the other three cases lately mentioned, however, there is a causal process connecting the events in question; so we should bite the bullet, and accept transitivity.

I reject causal transitivity for independent reasons. But I agree with Hall that double-prevention cases should be regarded with some suspicion: the events may be counterfactually dependent each on the previous, but it is not always natural to see them as a chain of causes and effects. However I reject Hall's diagnosis: in particular, I reject the suggestion that *causal process* has anything to do with the matter. This distinction between causal process and lack thereof is too important to leave undefined. If a causal process is simply a

chain of direct causation (cf. McDermott 1995) then whether there is a causal process will depend on whether there is a chain of direct causation. Hall is arguing that we must reject either the dependence thesis or transitivity; the absence of a causal process between the trip and the yelp is supposed to motivate rejecting the thesis that the dependence of the trip on the yelp suffices for the trip to cause the yelp. But on this definition of causal process, the argument is circular: the trip does not cause the yelp, because there is no causal process; but to say there is no causal process is just to say there is no chain of causes and effects from trip to yelp, which implies that the trip does not directly cause the yelp either (assuming, as seems reasonable, we allow one-step chains). Another definition of causal process might be offered, of course; until then, this line of argument is unlikely to sway a defender of the dependence thesis. And since this case forces us to give up either dependence or transitivity, this argument is not a solid defence of transitivity.

But I agree, and I think that the Reverse Counterfactual account agrees, that double-prevention is not causation. Let me now propose my own alternative diagnosis.

5.6 Refinements

5.6.1 Diagnosing Double-Prevention Counterexamples

There is potential for dialectical confusion here. I have argued that causation is not transitive. I have distinguished two sorts of cases where it appears to fail: distance and double-prevention. But now I am going to argue that double-prevention counterexamples do not really constitute counterexamples to causal transitivity at all, because they fail to amount to causal chains. Why, given that I reject transitivity, do I also reject these counterexamples to transitivity?

The reason is as follows. In a case of double-prevention, the only reason for thinking that the first preventer causes the second is that Lewis's counterfactual is true of it. But I deny that Lewis's counterfactual suffices for causation. Moreover, I adopt the Reverse Counterfactual as a necessary condition for causation; and in each of the cases we have considered, the first preventer fails to satisfy the Reverse Counterfactual with regard to the second. The only reason we accept claims like, "the massage was a cause of his going to New York", is that the massage is a *condition* for his going to New York. If we endorse a selective notion of causation, this reason is not a decisive one.

Let us consider each example, and argue that the first preventer fails to

cause the second by intuitive lights, when we are careful to distinguish causes from mere conditions. (I will assert the denial of the Reverse Counterfactual claim in brackets after each intuitive argument. Since I am denying that causation occurs in each case, I will be denying the Reverse Counterfactual in each case; and this is equivalent to asserting a might-counterfactual. So I will assert the appropriate might-counterfactual, which I think is easier to grasp.)

Consider the Skiing Accident. I have already argued, inspired by Anscombe, that the accident is not a cause of Suzy writing the paper, even if it does cause her to write it with her left hand. (If Suzy had not written the paper, she might still have had a skiing accident.) Similar remarks apply to Dog Bite: the dog was not responsible for McDermott's bombing, only for his choice of finger. (If McDermott had not pushed the button, the dog might still have bitten off his finger.) In the case of Shock C, where Able wants to save C and Baker wants to shock him, perhaps the most strained aspect of the example is that we are asked to see the well-meaning Able's move as the cause of Baker's. But arguably Able's move did not cause Baker's, even though it was a mere condition for it. (If Baker had not thrown the switch, Able might still have thrown his.) We would more naturally attribute Baker's action to his own malice, or his skill at the game, just as we attribute the goal to the striker and not to the goalkeeper — even though, if the goalkeeper had been positioned differently in the goal mouth, the striker would have gone for the other corner of the goal.

In each of these cases, the temptation to see c as a cause of d arises from the fact c is a condition for d. But if we distinguish causes from mere conditions, as I suggest we should, then it becomes less convincing that c causes d in each case. Similar remarks apply to the other cases. Billy could easily have run and not fallen; running may be a condition for the trip, but did not make it happen. (If Billy hadn't tripped, he still might have run.) And the assassin planting a bomb under Hall's desk is only seen as causing Hall to find it if we accept that counterfactual dependence suffices for causation, which I have argued we should not accept. (If Hall hadn't found the bomb, the assassin might still have planted it.) The chest massage may be a cause of the survival, but it is hardly a cause of the journey to New York. (If the victim hadn't gone to New York, he might still have had the chest massage.) And Kvart's accident is only seen as causing the surgery in a special sense; it is not as if the accident made the surgery happen — that was the job of the surgeon. (If Kvart had not had surgery, he might nevertheless have injured his finger.) Finally, the

generator fires up when the national grid goes down. The longer the generator runs, the harder it is to see the grid failure as causing the generator to run. It does not *make* the generator run. The only sense in which it might be seen as a cause is the sense I am arguing we should *distinguish* from causation: if the grid wasn't down, the generator wouldn't be running. That might be enough to explain why the generator is running, if someone asks, but it does not make the generator go in the way that diesel does. (If the generator had not come on, the grid might still have gone down.)

We might wonder whether it is a mere coincidence that each example shares the feature I have capitalised on. When c is the first preventer, d the second preventer and e the effect, is it always a strain to see c as causing d? The answer is, yes: in principle, cases of double prevention always have this feature. It is not a coincidence. In a case of double prevention, c is a mere condition for d, and d causes e. But it is always a strain to see c as causing d because, intuitively, d prevents c from causing e. In a case where c and d are suitably independent for the example to have intuitive force — where they are not parts of the same mechanism, or some such — we will usually be inclined to see the second preventer as caused by something appropriately independent. Cases where c and d are not independent — for example, where they are conflicting parts of a badly-built machine — we are not so likely to regard as failures of transitivity, but rather as complex causal processes.

Another consideration compels us to reject the claim that c causes d in cases like these. In this chapter, I have claimed that causation is neither transitive nor intransitive. It is possible for c to cause d and for d to cause e, but for c to fail to cause e. Since I have proposed the Reverse Counterfactual as a necessary condition for causation, this means I must allow that the following claims are both logically consistent and consistent with any further constraints on, or claims about, causation:

(1) $\sim D > \sim C$ [because c causes d]

(2) $\sim E > \sim D$ [because d causes e]

(3) $\sim (\sim E > \sim C)$ [because c does not cause e]

Inspired by the suggestion that nontransitivity be seen as a case of causal selection, we might be tempted to accept that c may also be a condition for d, and d for e, and indeed that c may be a condition for e — yet c still not cause e. That is, we might be inclined to accept the following Lewisian counterfactuals, in at least some cases (those not exhibiting redundancy):

(4) $\sim C >\sim D$ [because c is a condition for d]

(5) $\sim D >\sim E$ [because d is a condition for e]

(6) $\sim C >\sim E$ [because c is a condition of e, despite not causing e]

We might be inclined to accept this with regard to either kind of counterexample — proximity failure, or double prevention. So we might accept it of the chain birth – car crash – death, or we might accept it of the chain grid failure – generator starts – power supplied.

But we should not, because the combination is inconsistent. A valid substitute for transitivity is this :

$B > A$

$A > B$

$B > C$

$-------$

$A > C$

(Lewis 1973c, 17)

If we plug in (5), (2) and (1) in that order, we entail the negation of (3). That is:

(5) $\sim D >\sim E$

(2) $\sim E >\sim D$

(1) $\sim D >\sim C$

$-$ $-------$

$\sim$(3) $\sim E >\sim C$

Of course, (3) states the key claim $\sim (\sim E >\sim C)$: that the Reverse Counterfactual may be false of c and e even when they are related by a chain of counterfactual dependence (of any kind).

There is a more general question here, about how consecutive causal relations add up to one big one. For when they do, it is not an extra and unrelated fact that they do: when c causes e, it does so because it causes d, which causes e. But if causation is not transitive, what does it mean to say that these intermediate causal relations *give rise* or *add up* or *amount* to one big one?

Compare touching, an ordinary nontransitive relation. Jack is touching Jill and Jill is touching Jim; Jack might or might not be touching Jim, but if he is, it has nothing to do with whether either of them is touching Jill. Whereas in causal chains, the intermediate relations have everything to do with the overarching one. It might be possible for c to cause e without doing so via d, but then the case would be different. Whereas Jack and Jim might stand in just the same touching relation to each other, regardless of whether either was touching Jill. In other words, we need to explain why causation sometimes *seems* to be transitive, if it is not.

The kind of argument above offers a powerful potential explanation. For we could say that, although causation is not transitive, it effectively becomes so when redundancy is absent in the right places. In a chain of three events, each causing the next, the first will cause the third if the intermediate event is also a mere condition for the third. Note that this is a partial explanation: I do not say *only* if. There may be cases where the first event causes the third, but where this condition is not satisfied. (Most obviously, the cases of preemption which we will consider in Chapter 6 have this property.) The suggestion, rather, is that *when* this condition is satisfied, causation *must* be transitive. If we take longer chains, this generalises to the claim that causation must be transitive in causal chains where each cause is also a mere condition for its effect: that is, when there is no redundancy.

Although it is partial, this is quite a neat explanation, arguably more elegant and more discriminatory than just stipulating causation to be transitive across the board, as Lewis does. For every case where we accept that causation fails to be transitive, we must do one of the following:

- deny (1) that c causes d (so we may assert $\sim (\sim D > \sim C)$); or
- deny (2) that d causes e (so assert $\sim (\sim E > \sim D)$); or
- deny (5) that d is a condition for e (so assert $\sim (\sim D > \sim E)$).

I have just made an independent argument that we should deny (1) in cases of double prevention, on the basis that I recommend a selective notion of causation. The puzzle of double-prevention was that c may be a condition for d and d for e, yet c may fail to cause e. My solution, in a nutshell, is to assert that c is a mere condition, and not a cause, of d. The puzzle of double-prevention arises because philosophers often ignore this distinction (as illustrated by the venerable Mill-Lewis tradition). Moreover, causal terminology is just vague:

people can accept that the power failure caused the generator to come on, *in the sense that* if the power hadn't failed the generator wouldn't have come on. But this is a case of conversational charity; if we distinguish causation from conditionhood, we see that the first preventer is all condition, no cause, with regard to the second.

Now let us see if we can achieve a similar diagnosis of proximity failure.

5.6.2 Diagnosing Proximity Failure

Our discussion of the appearance of causal transitivity also explains why double-prevention counterexamples to transitivity exist. Does it explain distance counterexamples too? Not all distance counterexamples share the crucial failure of the first step in the proposed causal chain, which we used to explain double-prevention. For example, my birth is plausibly seen as both a cause and a condition of my first breath; and my first breath is a mere condition for my last; that gives us all the counterfactuals we need to prove that the Reverse Counterfactual holds between my last breath and my first.

But as we have already noted, claims like,

> If I hadn't been born, I wouldn't have just had a coffee

are ambiguous. On the one hand, they are obviously true, if the antecedent refers to the mere fact of my birth. But compare:

> If I hadn't been born, my grandparents would not have had a phone call at 3am that morning.

The same form of words in different contexts yields what are effectively two different antecedents. It is not plausible to suggest that my actual birth, specified in any detail, is a condition for many other events in my life, such as my going to university and studying philosophy. After all, many other people have had rather different births, at different times and in different places, with different parents and different degrees of ease or difficulty, yet some of them still seem to manage to study philosophy too. Returning to our previous example, my birth as it actually happened may be a both a cause and condition for my first breath; but the details of my first breath are irrelevant to the details of my last, unless they are very close together (which in my case they were not).

In sum, proximity failure occurs when the long chain of Lewisian counterfactuals fails. This is because, the longer the chain, the greater the possibility

of redundancy: the likelier we are to wonder whether the effect in question might not have come about by some other, maybe unguessed, route; so we stop regarding the earlier events as mere conditions for the later. This is the Paul-Anscombe point: it is hard to see the details of my coffee drinking as depending on the details of my birth. We can circumvent this difficulty by specifying events very blandly — the fact of birth, and the fact of drinking coffee. But then the causal links become less plausible. The mere fact of birth, considered on its own and without details, does not put a coffee in my hand, though it is a mere condition for my drinking coffee.

It would be nice to check the correctness of this analysis with a little thought experiment. Suppose someone provided a very long chain of events, and argued convincingly that each was a mere condition for the next, taking into account all the foregoing remarks about detail. Suppose it was also convincing that each event also caused the next. Then, if I am right, we would be inclined to accept that it also caused the last. In practice, however, it is extremely hard to do this, precisely because it is unconvincing to chain many events together and claim convincingly that each is a condition for the next. The more events we include, and the longer the chain gets, the easier it is to wonder whether one of these events might have come about anyway, by some other means. In such cases, we are not compelled to take the first event in the chain as the cause of the last. That, I think, is why causation fails to be transitive when causes are not proximate to their effects. I have not, however, shown positively that we should *not* take the first event as the cause of the last in such circumstances. Perhaps the reason for that comes down to further details of the selectivity of causation, which this account does not capture.

5.7 Summary

It has been argued that causation is not transitive. Despite the widely held view to the contrary, the four motivations we considered were found wanting. It is far from clear that the transitivity of causation is an "*a priori* insight" or a "bedrock datum"; at any rate, I do not share the insight. Lewis stipulates that causation is transitive to help his account handle preemption, but that way of dealing with preemption is the weakest aspect of Lewis's account (and no other account has gained general acceptance either). It also relies on the denial of backtrackers, a denial we have not endorsed. Both these points will receive further discussion in the next chapter. Lewis has an argument concerning the

actions of Apollo, but that argument was found wanting; moreover it was first published in 2000, so hardly explains the wide acceptance of transitivity for three decades or more previously. Opposition to a selective, or discriminatory or inegalitarian notion of causation was also considered as a motive for defending transitivity. It may be an important motivation, but this simply strengthens our case. We have already seen that there are good reasons to rebel against the anti-selective orthodoxy, making it all the more natural that we should question the doctrine of causal transitivity.

Two sorts of cases were distinguished in which causation fails to be transitive: cases where the cause is *too distant from*, or *fails to be sufficiently proximate to* the effect; and cases where there exists a special causal structure which I identified as *double prevention*. It was argued that the Reverse Counterfactual correctly reflects our intuitions about cases of proximity failure. At somewhat more length, the same was claimed in the case of double prevention, where we tested the Reverse Counterfactual against seven purported counterexamples to transitivity from the literature. In each case, the Reverse Counterfactual agreed that the first event failed to cause the last. Three defences of transitivity against these counterexamples were criticised, and their useful parts found to be either incorporated in or explained by the Reverse Counterfactual account. Finally it was noted that certain combinations of counterfactuals could not consistently be held.

This enabled us to deepen our understanding of causal nontransitivity. First, in a sense it corroborated the defences of transitivity in the literature, since we found reason to deny that the first preventer can legitimately be seen as causing the second, in a case of double prevention. I suspect the lack of a theoretical distinction between causes and mere conditions has led many thinkers to miss this feature of double-prevention cases. Second, again reducing the distance between my account and the common view, it was found that causation is transitive in any case where there is no redundancy and where each event genuinely causes the next. This partly explains why causation is so often treated as transitive over short chains, and also why the nontransitivity of causation fails to have the pure contingency of other nontransitive relations like touching. Finally, we were able to offer an account of proximity failure by noticing that the longer a causal chain, the less plausible it is simultaneously to maintain that each event strictly causes the next, and that it is counterfactually necessary for the next. The longer the chain, the greater the scope for redundancy, and for wondering whether later events might have been how

they actually were, even if earlier events had been a bit different.

Redundancy has played a big part in this chapter. Now let us give it our full attention.

Chapter 6

Redundancy[1]

6.0 Abstract

The term "redundant causation" is an umbrella. Overdetermination among existing events is distinguished from redundancy due to non-occurring back-ups. Symmetric and asymmetric overdetermination are further distinguished, and within the latter, preemption and trumping are distinguished. It is proposed that the Reverse Counterfactual holds true of cause-effect pairs even in cases of preemption. This gives us the basis for distinguishing causes from preempted events. I argue that the Reverse Counterfactual is true only of causes in some easy cases, allowing us to distinguish causes from preempted events. Then I argue that we can make the same distinction in harder cases, if we employ an intuitive notion of a causal chain. Next we discuss trumping, where appeal to causal chains appears not to help. I argue that the Reverse Counterfactual distinguishes trumping events from trumped events with whatever intuitive resources are provided to motivate the intuitive distinction; if no such resources are provided, trumping collapses into symmetric overdetermination. Finally we discuss symmetric overdetermination, which does not yield counterexamples in the way that preemption does. Nevertheless I argue that the Reverse Counterfactual account avoids the difficulties which beset Lewis's account associated with mereological summing.

6.1 Preemption, Necessity and Sufficiency

Before we dig into the main discussion, it might help to remember why various sorts of redundant causation — notably preemption — have caused problems

[1]This chapter is based on material in [Broadbent, 2007].

for counterfactual analyses. The heart of the problem is this. The tradition of counterfactual analysis starts with the thought that a certain counterfactual, $\sim C > \sim E$, is sufficient for causation (a claim which the last two chapters have disputed). The existence of redundant causation shows that this counterfactual cannot also be a necessary condition for causation. For in cases of redundant causation, it is false that $\sim C > \sim E$, yet true that c causes e.

The reason that redundant causation poses this sort of challenge to counterfactual accounts is as follows. Counterfactual accounts in the Lewis tradition make causes necessary, in a certain sense, for their effects. That is, the standard counterfactual account makes causes necessary for their effects in a counterfactual sense: in the sense that, had the cause not happened, the effect would not have happened. We might also say that the cause was necessary in the circumstances, or some such. But this is problematic: although causes are sometimes necessary in this way for their effects, in other circumstances they are not. There is often more than one way for a given effect to come about: often, causes are redundant in this sense.

At this point it is important not to get our necessities mixed up. Lewisian accounts offer a sufficient condition for causation which makes causes counterfactually necessary for their effects. By contrast, the Reverse Counterfactual offers a necesary condition for causation, which (I shall argue shortly) makes causes sufficient for their effects. In short:

Lewis's Account: If $\sim C > \sim E$ then c causes e.

Reverse Counterfactual Account: If c causes e then $\sim E > \sim C$.

Redundancy is what prevents Lewis's counterfactual being a necessary condition for causation as well as a sufficient one; and this is because causes fail to be counterfactually necessary for their effects in cases of redundancy.

Redundant causation is a problem for any account which makes causes necessary for their effects. An account which does not do this, will not suffer from redundancy counterexamples, such as cases of preemption. Giving an example of a case where c fails to be counterfactually necessary for e will be of little interest to an account which does not assert any such counterfactual necessity in the first place. Mine is such an account. The Reverse Counterfactual cannot be seen as asserting that causes are counterfactually necessary for their effects; in fact, it is better read as asserting that they are counterfactually sufficient (a claim I shall defend in a moment). That is to say, for c to cause e, the occurrence of c must be enough for the occurence of e, in a certain sense.

Prima facie, cases where c causes e yet where the occurrence of c fails to be required for the occurrence of e are irrelevant to this claim. The claim that c suffices for e is not in the least undermined by exhibiting cases where, if c hadn't happened, e might have happened anyway.

It is not uncommon to see writers generalising from the difficulties arising from redundant causation and suggesting that preemption is a problem for all counterfactual approaches. For example:

> ...the failed intermediary and would-be differences strategies that extant [counterfactual accounts of causation] use are inadequate as general solutions to the preemption problem. Of course there may yet be some new strategy for [counterfactual accounts of causation] which will prove adequate, but at this point the prospects look dark.
>
> (Schaffer 2004a, 71)

These sentiments echo Lewis's charismatic dismissal of the regularity approach, in favour of his promising new counterfactual alternative. But in the present context, such a discussion is hasty. The difficulties with preemption have nothing to do with counterfactual accounts *per se*. Rather, they arise because Lewis's is a necessity account of cause. A counterfactual account not sharing this feature need not suffer from preemption in the way that Lewis's account does.

The Reverse Counterfactual clearly does not make causes counterfactually necessary for their effects, but one might wonder whether it really makes them counterfactually sufficient. The most obvious way to characterise the notion that c is counterfactually sufficient for e would be to assert that, if c were to happen, then e would happen: that is, $C > E$. This is not equivalent to $\sim E > \sim C$, because contraposition fails for counterfactuals.[2] In general, the counterfactual necessity of A for B is not equivalent to the counterfactual sufficiency of B for A. And the Reverse Counterfactual appears to characterise the counterfactual necessity of the effect for the cause — a notion which has little intuitive appeal.

[2]Cf. Lewis 1973c, 17). For example: "If I had not been in the pub at 7pm yesterday, I would have been in the library" might be true, but "If I had not been in the library at 7pm yesterday, I would have been in the pub" might be false. The first counterfactual would have helped you find me, had you been looking, while the contrapositive would have led you astray. For as it happens I was in the library, not the pub, and if I hadn't been in the library, I would have been at home.

Nevertheless I think we should see the Reverse Counterfactual as characterising a kind of counterfactual sufficiency of cause for effect. On Lewis's semantics (and this is an aspect which I am not disputing), $C > E$ is true automatically when $C\&E$. So that is not available as a useful characterisation of the way causes are counterfactually sufficient for their effects: every pair of actual events would satisfy it. The Reverse Counterfactual might be seen as the nearest non-trivial alternative. More positively, the Reverse Counterfactual makes causes sufficient for their effects in a very clear way. For C and $\sim E > \sim C$ entails E, by counterfactual modus tollens. In other words, given that c causes e, the occurrence of c is logically sufficient for the occurrence of e. This is surely intuitively right: it is hard to see how c could happen, and cause e, and yet e fail to happen. For these reasons, then, I suggest that the Reverse Counterfactual be seen as a sufficiency account of causes; therefore we can be optimistic that it will cope well with the infamous preemption counterexamples which we are going to consider.

6.2 Varieties of Redundancy

Let us say that an event is *redundant* with respect to a given effect when it fails to be counterfactually necessary for it. The problem this poses for Lewis's counterfactual account of causation is then clear: sometimes, events which are redundant with respect to e nevertheless cause e. Among these, we can distinguish cases where c_1 causes e, but is redundant due to a non-occurring back-up event, c_2, which would occur and cause e if $c_1$did not. We can restrict the term *overdetermination* to cases where two or more actually occurring events are redundant with respect to a given effect, and e is caused by at least one of them. Then c_1 and c_2 both occur, but are redundant because if either were to fail to occur, e would still occur thanks to the other. Among these, we can further distinguish symmetric and asymmetric cases. In symmetric cases, c_1 and c_2 have equal claim to cause e; in asymmetric cases, it is intuitively clear that only c_1 and not c_2 causes e. Asymmetric cases may be subdivided yet further by distinguishing those where c_1 causes e in virtue of causing some intermediary event which causes e. Let us call these cases *preemption*. The remaining cases, asymmetric and without intermediary events, we call cases of *trumping*. Finally, for convenience, let us include cases of redundancy due to non-occurring back-ups with the cases of preemption. Figure 6.1 illustrates

Figure 6.1: Kinds of Redundancy

	Asymmetric cases	Symmetric cases
Non-occurring back-ups	Redundancy due to non-occurring back-up, which we shall include with our discussion of preemption	No cases, because the intuitive symmetry between two events is broken if one does not occur
Occurring back-ups	Preemption of an occurring backup, with failure of intermediary events	Symmetric overdetermination
	Trumping of an occurring backup, with no failure of intermediary events	

this classification.[3]

These distinctions differ somewhat from Lewis's. In particular, Lewis considers trumping to be a kind of preemption. I prefer to distinguish it, however, since I think it is a rather special sort of case (as I shall argue in 6.6). In addition, Lewis usually assumes that preemption occurs only between actually occurring events. However this distinction is not rigorously observed (eg. see Lewis 1986a, 195–6), and it will make our discussion neater if we include redundancy due to non-occurring back-ups with cases of preemption among actually occurring events.

Most of the discussion of preemption focuses on a distinction which I have not given: the distinction between early and late preemption. This distinction is not intuitive, and cannot only be given in terms of the first of Lewis's three solutions to the problem of preemption. Let us therefore introduce the first solution, and the rival here to be preferred.

[3]Throughout, when I speak of a non-occurring event I mean a possible event which does not actually occur. "Non-actual event" or "potential event" could equally have been used; I prefer "non-occurring" because it can be used also for the term "back-up", whereas talk of non-actual or potential back-ups might be confusing.

6.3 Early Preemption

6.3.1 Lewis's First Solution: Chains

Able and Baker are philosophical vandals. They enjoy throwing rocks at bottles and falsifying theories of causation. One day, they set up a bottle, retire to a respectable distance, and select some good rocks. Able throws a rock, and smashes the bottle. If Able hadn't thrown, Baker would have done; so the bottle might have smashed. So it is false that, if Able hadn't thrown, the bottle wouldn't have smashed. Yet Able's throw caused the bottle to smash.

Lewis's analysis starts with the thought that c causes e if $\sim C >\sim E$. Preemption counterexamples make it difficult to add a necessary condition to this sufficient condition. Lewis's first proposal is as follows. Causation is to be identified, not with counterfactual dependence, but with chains of causal dependence.[4] There is causal dependence between actual and distinct events c and e iff $\sim C >\sim E$ (Lewis 1986a, 166–7). c causes e iff there is a chain of causal dependence running from c to e.

There can be causation without causal dependence, because of the failure of transitivity for counterfactuals. Suppose e causally depends upon d and d upon c. Then c causes e. But it does not follow that e causally depends upon c: it is logically possible that, if c had not happened, e might still have happened (Lewis 1986a, 167).[5]

Lewis's solution to early preemption is bipartite, and employs this point first. Although the smash does not causally, or counterfactually, depend on Able's throw, there is a chain of dependence leading from the throw to the smash. The chain might go like this. Able throws, and if he hadn't thrown, his rock would not have hit the bottle — Step One. If his rock had not hit the bottle, the bottle wouldn't have broken — Step Two. Thus we have a chain of events, and each counterfactually depends on the previous, even though the last does not counterfactually depend on the first.

But might we not protest that, if Able's rock had not struck the bottle, maybe Able wouldn't have thrown, and then Baker would have thrown, and so the bottle might have broken anyway? (We would be objecting to Step Two in the chain above.)

No, says Lewis, because this objection is based on a piece of backtracking

[4] As previously noted, to claim causal dependence between two events is equivalent to claiming counterfactual dependence plus the claim that the two events actually occur (2).

[5] Lewis also protects the claimed transitivity of causation against the nontransitivity of counterfactuals with this point.

reasoning. It includes the claim, "If Able's rock had not struck the bottle, Able wouldn't have thrown", and that is clearly a backtracker. This invocation of the asymmetry of counterfactual dependence is the second part of his solution to early preemption.

> It remains only to reply to the objection that [the effect] e does *not* causally depend upon [the intermediate event] d, because if d had been absent then [the cause] c_1 would have been absent and [the preempted] c_2, no longer preempted, would have caused e. We may reply by denying the claim that if d had been absent then c_1 would have been absent... I rather claim that if d had been absent, c_1 would have somehow failed to cause d. But c_1 would still have been there to inferfere with c_2, so e would not have occurred.
>
> (Lewis 1986a, 172)

Lewis asserts that if Able's rock hadn't hit the bottle, then he would still have thrown, but for some reason the rock would not have hit the bottle.

Notice two things. First, Lewis does not merely deny this backtracker:

> If Able's rock hadn't hit the bottle, he wouldn't have thrown.

He does not merely deny $\sim D >\!\sim C_1$. Rather, he asserts $\sim D > C_1$: that is, he asserts:

> If Able's rock hadn't hit the bottle, he *would* still have thrown.

The reason is that merely denying the backtracker in question would amount to asserting $\sim (\sim D >\!\sim C_1)$, which is equivalent to $\sim D \geq C_1$. In words:

> If Able's rock hadn't hit the bottle, he *might* have thrown.

But obviously this is compatible with —

> If Able's rock hadn't hit the bottle, he might *not* have thrown.

And that is enough to run our objection: Able might not have thrown, so Baker might, so the bottle might have broken anyway. That is enough to falsify the claim that it would not have smashed, had Able's rock not struck it. To rule this out, Lewis must not merely deny the backtracking reasoning powering the objection; he must employ some of his own.

The obvious point here is that Lewis endorses a backtracker of his own. But I don't think this is a killer point: I think he would say this is one of the backtrackers licensed by his account (see 2.4.1). Perhaps less obvious but more important is the point that merely *denying* backtrackers is not enough to make Lewis's solution to early preemption work. Lewis's arguments concerning the asymmetry of counterfactual dependence support *denying* backtrackers, but nothing more. If Lewis's solution does indeed rely on the asymmetry of counterfactual dependence, then it does so in a more subtle and less obvious way than that which he presents.[6] I have not seen this point attended to.

Second, the notion of preemptive causes "interfering" with preempted causes is somewhat distant from our ordinary understanding of the term "preemption". On Lewis's picture, Able's throw preempts Baker's because it *prevents* it — at least, because it prevents it from causing a bottle smash, by preventing either the throw itself or some event further along the causal chain which would otherwise lead from Baker's throw to the smash. But this is not how we ordinarily use the term: usually, preemption involves no interference; the preemptor just gets there first. Lewis's slight deviance from ordinary usage suggests the obvious place to look for more difficult counterexamples, namely, in places where there is no interference between the two causal processes, and the preempting one indeed just gets there first. Late preemption cases are like this. We will come to the distinction between early and late preemption shortly, but first, let me advocate a different solution to cases of the sort we have just discussed, where redundancy is due to the presence of non-occurring back-ups.

6.3.2 The Reverse Counterfactual Solution

We saw in 6.1 that the accounts which make causes sufficient for their effects ought not, in principle, be threatened by preemption counterexamples. Cases of preemption show that causes are not always counterfactually necessary for their effects, thereby threatening the claim that counterfactual necessity is itself a necessary condition on causation. But a necessary condition on causation which claims that causes are always counterfactually *sufficient* for their effects will not be counterexampled when causes fail to be counterfactually necessary for their effects. Such are preemption cases, and the Reverse Counterfactual is a sufficiency account of cause; so it ought, in principle, not be threatened by preemption.

[6] Presumably by depending on the backtrackers which, as we have seen, his account implies.

And indeed, it seems that the Reverse Counterfactual does not need any qualification to deal with our example of preemption. The closest world in which the bottle does not smash is one where Able does not throw the rock, or misses. True, nor does Baker. But Baker does not actually throw a rock. So he can hardly be said to actually cause the smashing of the bottle: causes must be actually occurring events.

Maybe that was too quick. Is the nearest world where the bottle doesn't smash one where Able doesn't throw (or misses)? What justifies the claim that it is?

Suppose you were watching Able and Baker throwing rocks at the bottle. You watch Able take aim and then your attention is distracted for a moment. You look back, and see Able and Baker casting around for good-sized rocks, and the bottle still intact. "What happened, did you miss?" you call out. "Or perhaps you decided to look for a better rock, and didn't throw?" If you share the intuition that this is how you would respond, then I think you share the intuition that the closest world where the bottle does not smash is one where Able does not throw, or throws and misses. You have arrived at that conclusion by an application of the Inference Test: you supposed the antecedent true (that the bottle didn't smash), and inferred the consequent (that Able didn't throw accurately, or at all).

It will be noticed that I have inserted the qualifier "accurately". It might be objected that the closest worlds where the bottle doesn't smash include some where Able misses, and thus it is false that if the bottle hadn't smashed, he wouldn't have thrown (because he would have thrown and missed). Actually I think this objection brings out an advantage of the account. Accuracy is causally relevant, in this case. The cause of the smash is Able's accurate throw. The fact that the Reverse Counterfactual specifies that Able throws accurately (by specifying that if the bottle hadn't smashed, Able wouldn't have thrown or would have missed) is hardly a criticism.

It appears, therefore, that reversing the Lewisian counterfactual provides us with a counterfactual which can deal with this example of preemption with no modification at all. This is very promising. Note, however, that this case of preemption has a special feature. The preempted event does not actually occur: Baker does not actually throw. The distinction which Lewis draws between early and late preemption does not correspond to a distinction between cases where the preempted event does not occur, and those where it does: early

preemption includes cases where both candidate causes actually occur.[7] So the Reverse Counterfactual solution is not really a solution to early preemption, as Lewis defines it, but to a limited set of cases where the preempted alternative does not actually occur. Nevertheless I shall argue that the solution extends to cases where both causes do occur, whether they are cases of early or late preemption.

Note that, even though Lewis requires causes to be actual events, our move is not available to Lewis. His is the problem that $\sim C_1 > \sim E$ is false, even if the competing c_2 does not occur. Whereas the Reverse Counterfactual faces a charge, not of falsity, but of too much truth. This is no problem for a necessary condition. Another necessary condition for causation, the actual occurrence of causes, can then be appealed to, in order to rule the preempted non-actual event out.

6.4 Late Preemption

6.4.1 The Early/Late Distinction

Now that they have limbered up, Able and Baker start a quick-fire bottle smashing competition. Both throw as quickly as they can at a given signal. Able's rock breaks the bottle, and moments later, Baker's rock whistles through the scattering shards where the bottle till lately stood. As in the previous example, if Able had not thrown, the bottle might still have broken. But Lewis's solution will not work. For unlike in our previous case, there is no chain of true Lewisian counterfactuals with which he can defensibly replace the single false one. Consider the impact event, Able's rock hitting the bottle. Suppose that hadn't happened, by some miracle: the rock swerves slightly. Baker's rock was just a few centimetres away; surely, it would — or at least might — still hit the bottle. And might is enough: if the bottle might still have smashed, then it is false that it would not have. This reasoning does not rely on backtracking of any kind, so Lewis cannot avail himself of his asymmetry of counterfactual dependence to rule it out of order. When we imagine that Able misses, we need not reason back to the absence of any earlier event, because Baker's rock is already on its way.

Although it is a difference between our two examples, the difference between early and late preemption in general does not come down to the difference

[7] And, I shall suggest, if we allow preemption of non-occurring back-ups, then these can be cases of late-preemption.

between cases where a preempted event occurs and those where it does not. So what exactly is the distinction between early and late preemption?

> In early preemption, the process running from the preempted alternative is cut off well before the main process running from the preempting cause has gone to completion. Then somewhere along that main process, not too early and not too late, we can find an intermediate event to complete a causal chain in two steps from the preempting cause to the final effect. The effect depends on the intermediate, which depends in turn on the preempting cause.
>
> (Lewis 1986a, 200)

The distinction is thus theory laden, depending upon the circumstances in which Lewis's first solution to the problem of preemption fails. It is, of course, possible to specify these circumstances, as Lewis does above. First, at some point in the preempted causal chain, there is a non-actual event which would occur if some event in the actual chain did not. Second, there must be at least two further events between the preventing event in the actual chain, and the effect (cf. Lewis 1986a, 200). These circumstances may be described, but I do not see any special significance about them. (Nor do I see Lewis claim any further significance for them.) I draw attention to the fact because the Reverse Counterfactual does not respect the early/late distinction, as we shall see. If the distinction is a theoretical one, and does not have a clear intuitive basis, then this need not worry us in itself. Nevertheless the early/late distinction is a big deal for Lewis, precisely because it appears to be a distinction between cases he can handle by denying backtrackers in the way described, and cases he cannot handle in that way.

Late preemption has proved to be the thorn in the side of counterfactual analysis. It would be a serious task to survey all the responses that have been made, and also a dubious expenditure of energy, since none has won a widespread following. Instead we will briefly consider Lewis's own two subsequent efforts to handle preemption.

6.4.2 Lewis's Second Solution: Quasi-Dependence

Lewis provides a tentative solution in terms of *quasi-dependence*, which we will briefly consider. His later solution in terms of influence is arguably more important, however, since it constitutes a thorough revision of his theory of causation.

Quasi-dependence between events consists in their being exactly like a nomically possible set of events which display counterfactual dependence. The thought is intuitively grasped. Lewis thinks that causation is an intrinsic matter (except, perhaps, for the involvement of the laws of nature); if there are two spatio-temporal regions governed by the same laws, which match exactly in matters of intrinsic fact, then there cannot be causation in one case but not in the other (Lewis 1986a, 206).

The intuitive observation underlying this account is that cases of preemption include events which are not causally relevant, but which affect the truth-value of the counterfactuals. Baker's throw doesn't have any effect on the bottle, it just changes the truth-value of the counterfactuals concerning what would have happened to the bottle if Baker hadn't thrown. We can imagine Baker's throw away without changing the causal facts concerning Able and the bottle. In effect, Lewis's quasi-dependence account suggests doing exactly this. There is a similar possible situation in which Able's throw and the bottle smash are exactly as they are, but Baker's throw is absent; and in *that* situation, the bottle smash counterfactually depends on Able's throw. In virtue of this, it quasi-depends on Able's throw in the actual world.

Lewis did not endorse quasi-dependence eagerly.

> The extended analysis, which allows causation by quasi-dependence, is more complicated than my original analysis... While I would still welcome a different solution to the problem of late preemption, within my original analysis, I now think that the extended analysis [in terms of quasi-dependence] may well be preferable.
>
> (Lewis 1986a, 207)

Not only was Lewis tentative about it in the first place, he produced a superseding theory, which we shall shortly discuss. Several obvious difficulties beset quasi-dependence. First, it sits uneasily with Lewis's (suitably qualified) view that causation is intrinsic. On a quasi-dependence account, what makes a causal process causal its similarity to some other possible process. Similarity to some other process is not an intrinsic property. This is particularly awkward given the role that intuitions about the intrinsicness of causation are supposed to play in motivating quasi-dependence in the first place. Second, it makes nothing of Lewis's strong intuition that *event fragility* has something to do with preemption. This is rather a therapeutic worry; nevertheless it could well be something preventing Lewis's fuller endorsement.

Finally, the position does not really seem to capture whatever features of our concept of causation power our intuitions about preemption cases. We could accept the thought that two intrinsically alike processes must be causally alike, but why should we say that causation occurs in both if counterfactual dependence occurs in one? We could equally deny causation of both, as far as the directive goes, that intrinsically alike processes are causally alike. Yet to do so would be ridiculous, in cases where causation is clearly present in both. Moreover it is arguable that our intuitions about causation in preemptive cases do not derive from their possible likeness to non-preemptive cases. We might accept that, for any case of causal redundancy, there is a possible case that is intrinsically similar with respect to the cause and the effect, but which displays no redundancy. But we might expect that this should be explained *by* the fact that the process is causal in the preemptive case, rather than vice versa.

6.4.3 Lewis's Third Solution: Influence

Lewis's most recent efforts to deal with late preemption (along with trumping, which we will discuss next) involve a significant adjustment of his view. In place of counterfactual dependence characterised straightforwardly by counterfactual conditionals, he introduces a notion of *influence*. The idea is to make the effect depend counterfactually not just on whether the cause happens, but on when and how it happens. For some putative cause c (Able throwing a rock, say) and some putative effect e (the bottle smashing), there is a range of alterations, $c_1, ..., c_n$, and $e_1, ..., e_n$. These are events each occupying a different world, and each differing from c and e to a greater or lesser degree. Strictly, c and e count as alterations of themselves. We don't need to decide at what degree the differences threaten the identity of the events in question — at what point the alterations stop being alterations to c and e and start being different events. This will be a strength of the position, if we agree with Lewis that our intuitions regarding event identity are frequently indeterminate and fickle (Lewis 2004a, 86). Now we can say:

> Where C and E are distinct actual events, let us say that C influences E iff there is a substantial range $C_1, C_2, ...$ of different not-too-distant alterations of C (including the actual alteration of C) and there is a range $E_1, E_2, ...$ of alterations of E, at least some of which differ, such that if C_1 had occurred, E_1 would have occurred, and if C_2 had occurred, E_2 would have occurred, and so

> on... C causes E iff there is a chain of stepwise influence from C to E.
>
> (Lewis 2004a, 91)

The intuitive idea is easy enough to see. If Able had thrown a little harder, or chosen a different rock, or aimed at the neck rather than the base of the bottle, the bottle would have smashed a little differently. Whereas Baker's whims would have made no difference.

Causation as influence has had a mixed reception in the literature. On the one hand, it overcomes the problems recently sketched for quasi-dependence. It does not generate tensions about the intrinsicness of causation, it explains why making events extremely fragile seems to help with preemption, and it explains why our intuitions about causation in preemptive cases do not seem to derive from our intuitions about non-preemptive cases. On the other hand, it lacks the striking simplicity of Lewis's original proposal. More worryingly, it has been called a subject change: causal influence and causation may turn out to be two distinct concepts, both of which we possess. They may even be extensionally distinct concepts. Collins cites one of Lewis's own earlier examples:

> If a poison kills its victim more slowly and painfully when taken on a full stomach, then the victim's eating pudding before he drinks the poisoned potion has a causal influence on his death, since the time and manner of the death depend counterfactually on the eating of the pudding. Yet the eating of the pudding is not a cause of his death.
>
> (Collins 2004, 114)

Substituting causal influence for causation risks counting all sorts of things as causes which we would not intuitively admit, although they make a difference to how the effect happens. If the light breeze that was blowing when Able threw had been a little different, then the bottle would have smashed a little differently; but our common judgement does not allow that the breeze caused the smash. I criticised Lewis's simplest counterfactual account for being unselective, but this amendment seems to make the problem a great deal worse. Even those who see no problem with an unselective notion of causation might baulk at accepting that the breeze which ruffled my hair on the way to the office is a cause of my arrival in the office, that arrival being — as it

was — a ruffled one, with many differently ruffled alterations corresponding to different possible breezes. More generally, universal gravitation means every massive object influences every other: a range of nearby alterations exist in which I move my hand more or less to the left, corresponding to each of which is a minute alteration of Saturn's orbit. But I do not claim to cause Saturn's heavenly progress just by waving my hands around, even though I exert an influence on the orbit.[8]

Another problem is that some sorts of cause-effect pairs may not exhibit the sort of variation Lewis envisages. This will particularly be the case when the effect is an all-or-nothing sort of event, either by nature or by circumstance. Crediting a conversation with Yablo, Hall points out that Able's rock could be replaced by a Smart Rock — a computerised jet-propelled rock, which can be programmed to attain a particular velocity and orientation when it hits the bottle (Hall 2004b, 237). Minor variations in Able's throw now count for nothing (though we might say he cheated in the quick-draw bottle smashing contest).

When Lewis discusses the question of whether making events fragile would rule out all cases of preemption, in the postscripts to his 1973 paper, he says this:

> ...residual cases of redundancy, in which it makes absolutely no difference to the effect whether both causes occur or only one... probably... would be mere possibilities, far-fetched and contrary to the ways of the world. Then we could happily leave them as spoils to the victor.
>
> (Lewis 1986a, 197)

In this vein, perhaps it could be argued that invoking computerised guidance systems for rocks is a bit far-fetched? — But there is no need to go hi-tech. My alarm clock is set to go off at 7am if someone presses the *On* button any time in the previous twenty-four hours. When I go to bed, I press it; but had I bothered to look I would have seen that the alarm was already set, because my wife had pressed it earlier. My wife's pressing the *On* button is the cause of the alarm going off, but if she hadn't pressed it, it would still have gone off because I pressed the button too. Moreover, the alarm's going off at 7am the next morning is highly insensitive to minor variations in how and when

[8] I owe the point about gravity to Dan Heard.

my wife pressed the button: its time and manner are not affected by small differences in the time and manner of my wife's action.

I do not propose to go any further into the details of Lewis's analysis, nor the possible problems it might face. Nor do I propose to discuss any other attempts to resolve the problem of late preemption. Instead, I would like to illustrate how these kinds of problems are very readily avoided by the simple reversing move I have been advocating, without appeal to quasi-dependence, causal influence, or any resource beyond the simple and intuitive notion of a causal chain.

6.5 Developing The Reverse Counterfactual Solution

6.5.1 The Modified Condition

Able and Baker both throw, and Able's rock gets there moments before Baker's. If the bottle hadn't smashed, neither Able nor Baker would have thrown accurately. The Inference Test yields this result: we would surely infer the absence or inaccuracy of both throws from the survival of the bottle. In worlds talk, at the nearest world where the bottle doesn't smash, surely neither Able nor Baker throws accurately. These are both salient differences from the actual world. The Reverse Counterfactual is true of them both. But in this competitive context, Able will insist — rightly — that his throw and not Baker's caused the smash.

There are two (compatible) ways we could respond, one more boring and one less. The more boring way would be to point out that the Reverse Counterfactual is offered as a necessary condition for causation: pointing out that it is met by non-causes does not threaten the claim that all causes meet it. However, this response is boring. Perhaps it would not be entirely uninteresting to discover that the Reverse Counterfactual is true when c_1 causes e, even if it is also true of the preempted c_2. But we might have hoped for more: just as we used the necessary condition consisting in the Reverse Counterfactual to discriminate between causes and mere conditions, we might have hoped to use it to discriminate between causes and preempted events. This is the less boring response, and the one which we shall explore here.

Notice how the problem for the Reverse Counterfactual is different from the problem Lewis's approach faces. The problem for Lewis is the *falsity* of

$\sim C >\sim E$, and something needs to be found to replace it — a chain of causal dependence, quasi-dependence, influence or one of the many more or less exotic theories which we have not considered here. The problem with the Reverse Counterfactual, however, is not falsity, but *too much truth*: it is true of causes *and* of preempted events. Seen the less boring way, this is a challenge to the discriminatory power of the proposed necessary condition for causation. To meet this challenge, we need to find some further feature distinguishing cause from preempted event, given that the Reverse Counterfactual is true of both. Moreover, as noted previously, we need this solution not just for late preemption, but for any case of preemption where both preempting and preempted events actually occur: for the previous solution relied on the non-occurrence of the preempted cause (the fact that Baker didn't throw in that case).

To find this distinguishing feature, imagine for a moment that we were asking, not about Able and Baker's throws, but about the *impact* of Able's rock on the bottle. We have two rocks, A-rock and B-rock. A-rock hits the bottle, causing it to break. But A-rock is redundant, because if A-rock hadn't hit the bottle, B-rock would have, and the bottle would still have broken.[9] Since this case is just like the easy case where Able threw and Baker didn't, it is predictable that the Reverse Counterfactual handles it easily. If the bottle hadn't broken, A-rock wouldn't have hit it: nor indeed would B-rock have; but B-rock does not actually hit the bottle, so B-rock hitting the bottle can hardly be said to cause the break. The Reverse Counterfactual is one necessary condition on causation, and the actual occurrence of causes is another: and only the impact of A-rock meets both; the preempted impact of B-rock satisfies the Reverse Counterfactual, but fails to occur.

But if we can distinguish A-rock and B-rock in this way, then surely we can distinguish Able's throw from Baker's. Lewis's original solution fails to distinguish A-rock from B-rock. And the failure to distinguish A-rock from B-rock is the reason — the whole reason — why Lewis's first solution fails for late-preemption; for it is at this point in the chain that we run out of true counterfactuals.[10] But the Reverse Counterfactual can distinguish A-rock from

[9]This is a case of late preemption, because there is no suitable event in between which Lewis might use to deploy his first solution. It is also a case, we can suppose, where there is no chain of events connecting cause to effect. This shows again that the early/late distinction fails to match the occurring/non-occurring preempted event distinction. There are thus cases of late preemption of non-occurring back-ups, as well as cases of early preemption of occurring back-ups (which are the ones Lewis focuses on).

[10]There is (we can suppose) no chain of events connecting A-rock's strike with the bottle's

B-rock; so one would think it ought to be able to distinguish between Able's throw and Baker's throw.

Let us ask the obvious question. Why do we think that Able's throw caused the bottle to break, but that Baker's did not? The answer is equally obvious: because Able's rock hit the bottle, whereas Baker's did not. This is not an accident of the example. The way we have defined preemption, there is never a causal chain from a preempted event all the way to the effect. And this, in turn, is no artifice of definition, but a highly intuitive description of the obvious reason we give for saying that Able's throw caused the smash, but Baker's didn't.

An obvious suggestion, then, is that for c to cause e, there must be a chain of events from c to e, each causing the next. But this obvious suggestion is wrong, because events can also cause each other directly, without intervening chains. (Indeed the events in the chain must cause each other directly, unless we are to posit an infinity of links in every causal chain.) The necessary condition which Able's throw meets and Baker's throw fails to meet in the bottle smash case is not that there is an intermediate event, but that *if* there is a chain of intermediate events, then it is unbroken: it goes all the way, as it were, to the effect. Able's throw satisfies this requirement, but Baker's does not. This has nothing to do with the Reverse Counterfactual *per se*; rather, it is a further necessary condition on causes which preempted events fail to satisfy. This is in line with the strategy we employed in the easy case, where Baker did not actually throw: there, we appealed to a platitudinous condition that causes occur, in order to rule out the competitors of which the Reverse Counterfactual was also true. The requirement that causes be connected to their effects by an unbroken chain is a less straightforward notion, and accordingly I now need to specify it more clearly.

I think the easiest, though perhaps not the only, way to tell whether a chain is unbroken is to see whether it changes under certain counterfactual suppositions. By definition, in a case where some candidate c is redundant with respect to e, there will always be further redundant events, which are counterfactually independent of c. Suppose those events away. Does the chain between c and e change at all? If so, you can rule c out as a cause.

Take Able's throw. There is a chain of Reverse Counterfactuals between the

---

subsequent shattering. The one causes the other immediately (at least, close enough for the everyday purposes which our everyday concept of causation serves). If A-rock had not hit the bottle, it would nevertheless have smashed, thanks to B-rock; and we don't need to employ any backtracking reasoning to arrive at this conclusion.

shattering of the bottle and the throw. But the throw is redundant; meaning nothing more than that it fails to be a condition for the smash — it fails to satisfy Lewis's counterfactual. Now suppose away any other events which are counterfactually related to the smash, but not to Able's throw, *and* which are also redundant. So we leave Able's chain intact: he throws, his rock flies through the air and strikes the bottle, and the bottle smashes. But we remove Baker's throw, because it is counterfactually related to the smash (by the Reverse Counterfactual — we can assume this, because it is the problem) and also redundant. (Grant that we have enough of a grasp of the causal order to be able to ignore events after the shattering, that is, effects of the shattering. The causal order will receive more discussion in the next chapter.) We can likewise suppose away the events in Baker's chain, such as the flight of his rock through the air towards the bottle. Now ask: Does anything change in Able's chain? Clearly not. Under this counterfactual supposition, a chain of Reverse Counterfactuals connects the same events in the same way from Able's throw to the smash. We can pronounce Able's chain *unbroken*.

Now hold Baker's throw constant, and remove other redundant events which are counterfactually related to the effect (by the Reverse Counterfactual). That means getting rid of Able's throw and the events in the chain leading to the shattering. It also means keeping the flight of Baker's throw towards the bottle, since that is related by the Reverse Counterfactual to the shattering on one side and to Baker's throw on the other. (If the bottle hadn't smashed, surely no rocks — Baker's included — would have flown towards it; and if Baker's rock hadn't flown, Baker wouldn't have thrown.) But there is another event which, under this counterfactual supposition, forces its way into the chain. In the absence of Able's throw, Baker's rock would have hit the bottle. And it would clearly have been related to the subsequent shattering in just the way that the actual impact of Able's rock is — meaning, I claim, that the Reverse Counterfactual would then relate it to the shattering on one side and to other events in Baker's chain on the other (Baker's throw, the flight of his rock). Under this counterfactual supposition, then, the chain of Reverse Counterfactuals between Baker's throw and the smash changes. This, I say, means we can pronounce the chain in the actual world *broken*.

I said that the unbroken chain condition was a further necessary condition on causation, in addition to the Reverse Counterfactual. It might then be doubted whether the Reverse Counterfactual account can really take credit for any success which this solution might enjoy. In response, note two points.

First, to reiterate, the problem which I face is one of too much truth: I have a necessary condition which non-causes meet, which is just to say that it is not a sufficient condition. Given that I claim only that the Reverse Counterfactual is necessary for causation, it is completely acceptable to point out other necessary conditions which causes meet and non-causes fail. Second, the Reverse Counterfactual is essential to the working of the distinction I have proposed between broken and unbroken chains. For I propose we suppose away all those redundant events which are nevertheless counterfactually related to the effect, except those chained up to the candidate cause under consideration. Since "redundant" means nothing more than "fails the Lewisian counterfactual", it is the truth of the Reverse Counterfactual which limits this supposition to redundant causes or their back-ups, and prevents it from extending to irrelevant events of all kinds. It is likewise the Reverse Counterfactual which forms the chain between the candidate cause under consideration and the effect. The distinction I have proposed between broken and unbroken chains therefore depends essentially on the Reverse Counterfactual.

I think this broken/unbroken chain distinction is an intuitive diagnosis of the preemptor/preempted distinction, and I think it generalises. In general, when c_1 preempts c_2 with respect to e, that is because c_1 causes d_1 which causes e. By contrast, c_2 does not cause d_1; and if c_1 were absent, then c_2 would cause some other event d_2, which would be a cause of e. But this event d_2 is preempted by d_1; and moreover, d_2 is preempted in such a way that it does not actually occur. The modification I propose for my account, then, is nothing more than a method for detecting events like d_2 — events which do not occur but which, were they to occur, would complete a causal chain between preempted causes and their would-be effects. I suggest that we can detect the place of an event like d_2 in the chain originating with c_2 by supposing away other redundant candidate causes.

When I say this is an intuitive diagnosis, I believe I am being honest. Commonly, we would say that Able's throw causes the break because his rock hits the bottle, whereas Baker's fails to cause the break because his rock does not hit the bottle. Perhaps this intuitive analysis is not entirely uncontroversial. It differs from the intuitive thought underlying Lewis's influence account, for example, which was that differences in Able's throw are reflected in the bottle smash whereas differences in Baker's are not. But we already saw some good reasons to doubt that influence is either necessary or sufficient for causation.[11]

[11]Smart Rock and the alarm clock show it isn't necessary; the breeze on the rock, the

Moreover, an intuitive initial position would have it that causes influence their effects (when they do) *because* they cause them, not vice versa. On the other hand, the intuitive analysis I have proposed is very close to Lewis's first thought — his solution to early preemption, the underlying idea of which was that there is a causal chain between cause and effect, but not between preempted event and effect. In this sense, we can agree with the general form of Lewis's original solution to preemption, although the details are quite different.

Let us summarise the foregoing discussion.

Reverse Counterfactual Modified for Preemption. If c causes e, then:

(i) $\sim E >\sim C$ [the Reverse Counterfactual requirement] and $C\&E$ [the occurrence requirement];

(ii) the chain of events $\{x, ...y\}$ between c and e such that $\sim E >\sim Y, ... \sim X >\sim C$ must be *unbroken* (where "between" is to be understood as including c and e in the limit case).

The point of the slightly deviant understanding of "between" is to allow counterfactuals with no events intermediate between antecedent event and consequent event to be subject to the test. We thus do not rule out immediate causation because $\sim E >\sim C$ counts as a chain, and hence could qualify as an unbroken chain if it passed the test. We can say define a broken chain as follows:

Broken Chains. A chain of events $\{x, ...y\}$ between c and e such that $\sim E >\sim Y, ... \sim X >\sim C$ is broken iff:

[redundancy] c is redundant with respect to e such that $\sim (\sim C >\sim E)$; and

[supposition] supposing away redundant events (other than e) which are counterfactually independent of c yet counterfactually related to e —

[new chain] the chain of events between c and e such that $\sim E >\sim Y, ... \sim X >\sim C$ under this supposition would differ from $\{x, ...y\}$.

This is the outline of my proposed solution. Now we shall turn to the defence of the key claims upon which it rests.

---

poisoned pudding and gravity show it is not sufficient.

6.5.2 Questions

Probably the clearest way to defend this proposal is to distinguish the questions we might ask about it, then answer them.

Is It True That $\sim D > \sim C$?

What is the basis for the claim that the Reverse Counterfactual holds between the actual impact d of Able's rock with the bottle, and Able's throw c? What is the basis for the claim that, if A-rock hadn't hit the bottle, Able wouldn't have thrown? Applying the Inference Test: would we infer that Able hadn't thrown, if his rock hadn't hit the bottle?

There are two ways to go. One way says that, if his rock hadn't hit the bottle, we wouldn't know what to infer. Maybe Able didn't throw: but maybe he missed, maybe he threw slightly slower, or maybe Baker threw slightly quicker. So the Reverse Counterfactual is false. The other way sets all these possibilities aside, and agrees that we would infer that Able hadn't thrown, and therefore that the Reverse Counterfactual is true.

Why should we set those possibilities aside? Well, we should not set them *all* aside. In line with my argument in 4.4.2, I suggest that the possibility that Able misses is relevant. All this shows is that the accuracy of Able's throw is causally relevant. If Able's rock had not hit the bottle, he would have either missed or not thrown. I am happy to accept that accuracy is part of the cause; nor do I have any objection to saying that the impact is jointly caused by his throwing and his not-missing, if it is insisted that these are distinct events. A similar response might be made to the suggestion that Able could have thrown a bit slower. In this context, the timing of Able's throw was crucial to his rock hitting the bottle: had he delayed, Baker would have got there first, and there would be no bottle for Able's rock to hit. Sometimes, timing matters: the context in which Able's throw occurs is one such occasion.

More difficult is the question whether Baker might not have thrown a bit quicker. This would have led to Baker's rock smashing the bottle, and thus Able's rock would not have hit the bottle; if such worlds are close enough, then it seems that if Able's rock hadn't hit the bottle, he might still have thrown. In a competition between two evenly-matched competitors, that seems just as likely as Able throwing a bit slower, or missing, or not throwing. And it is possible to devise set-ups which make Able's throw much less potentially variable than Baker's. Replace Able with a tennis-ball launcher, for example; Baker is practising his quick-fire bottle-smashing against a machine which

always launches a ball exactly 0.3 seconds after Baker's coach shouts "Go!" and presses the launch button. Surely Baker is more variable than the machine; yet it is still the firing of the machine which is responsible for the tennis ball hitting the bottle.

This is a version of Lipton's objection concerning background radiation which we considered in 4.6.2. The general stategy of this kind of objection is to devise a case where the cause is counterfactually rather stable, then argue that the nearest worlds where the effect does not occur will be ones where the cause still does occur. Thus in the present case, Baker's throw is more prone to variation than the tennis-ball launcher's launch, so if the effect had not occurred, it seems more of a departure to suppose that the launcher had not launched, than to suppose that it did still launch but that Baker threw quicker. (At any rate, it is not obviously *less* of a departure to suppose the latter.)

However, parallel to my argument in the case of background radiation, there may be contexts of inquiry where we hold Baker's throw constant and consider the ball-launcher the cause of the strike, even if the ball-launcher is less prone to variation than Baker. We might do so with a contrastive why-question: "Why did the tennis ball hit the bottle, rather than staying in the launcher?" More generally, the suggestion is this: in the context where we say that the launch c was the cause of the ball striking the bottle d, we say the counterfactual $\sim D >\sim C$ is true.

There are contexts where $\sim D >\sim C$ is false: in such contexts we say that if the ball had not hit the bottle, maybe the launcher would not have launched, but maybe Baker would have thrown faster. I suggest, however, that this is a good time to remember that our attribution of causation in cases of causal redundancy also display a certain sensitivity to context. We might equally say, in some contexts, that the reason the ball hit the bottle was that Baker's threw too slowly to smash the bottle before the ball got there. (We might be particularly so inclined if Baker usually beats the launcher: and if anything important hung on it, Baker might blame himself.) In this sense, we sometimes do admit that Baker's late throw was at least part of the cause of the smash — perhaps jointly with the launcher. Cases of preemption test our intuitions (contrary to the clarity that is often claimed for them). A revolutionary facing a firing squad is killed by a bullet fired by a single soldier, before any of the other bullets hit him; nevertheless, part of the purpose of a firing squad is to spread the responsibility for the execution over a larger

number of people. The redundancy in the case gives us a choice. In certain contexts, it is appropriate to say that the soldier killed the revolutionary, but in other contexts the massive redundancy makes it inappropriate: the revolutionary faced certain death even if the soldier had not fired.

Returning to the original example, where Able and Baker both throw, consider this question: "Why did Able's rock rather than Baker's hit the bottle?" We naturally respond, "Because Able threw before Baker (or perhaps he threw slightly harder)." We mention *both* their throws. In this context, then, it looks like we naturally consider Baker's throw to be part of the cause of Able's rock hitting the bottle. This should do something to upset the ease with which it is customary to say that, obviously, Able's throw causes the bottle to smash. Proper attention to our real intuitions makes this less than obvious. What is meant, and acquiesed to, by the claim that Able's throw "obviously" causes his rock to hit the bottle, is arguably something slightly less general than that bald claim. It might be the answer to a question like this: "What causes Able's rock to hit the bottle, rather than stay in Able's hand, or on the ground at his feet?" The answer to that question evidently has nothing to do with Baker's throw. But if we specify the context of inquiry that way, then the Reverse Counterfactual is satisfied by Able's throw and not by Baker's. The difference between the case where Able's rock hits the bottle and the counterfactual scenario where it stays in his hand, is Able's throw. Admitting that is tantamount to admitting that, if the rock hadn't hit the bottle but had stayed in Able's hand, he might still have thrown. Whereas Baker's throw is not a difference between these two cases: admitting which is tantamount to denying that, if Able's rock had not hit the bottle, then Baker would not have thrown.

In short, then, it is questionable whether $\sim D > \sim C$ is true in an unqualifed sense, but that is because it is questionable whether preemptors are intuitively seen as the sole causes of their effects. However, to the extent and in the context that we accept that a preemptor does solely cause its effect, it will be the case that $\sim D > \sim C$, i.e. that the Reverse Counterfactual is satisfied for the intermediary event d and the preemptive cause c.

Is It True That $\sim E > \sim D$?

Yes — I take it that this is the most obvious part of the proposal, and it has already received some defence in 6.5.1. It is easily seen with the Inference Test. If the bottle hadn't smashed, then why would we infer that A-rock — or any other rock for that matter — might have hit it anyway? Of course

it is quite possible that the rock would have just brushed it; but I think we can safely stipulate that that would qualify as a different event.[12] What this shows, again, is that mere contact did not break the bottle; the force of the impact was causally relevant. There are more outlandish possibilities — that the bottle was made of iron, that it had an internal bracing structure, that the rock was made of jelly, that a miracle occurred, and so on. But we would not infer any of these without a good reason, because they are outlandish. In worlds talk, they happen at more distant worlds; the nearest worlds where the bottle doesn't break are surely those where it is simply lucky enough not to get hit by any rocks.

Are We Relying On A Special View of Event Identity?

We might wonder whether we are using some special view of event identity. In particular, a distinction is made between the impact of Able's rock and the non-occurring impact of Baker's rock. If Able's rock had not hit the bottle, then Able wouldn't have thrown. But instead of speaking of Able's rock hitting the bottle, we might speak of an impact more generally; and had *no* impact occurred, then presumably neither Able nor Baker would have thrown. Then the Reverse Counterfactual does not distinguish Able's throw from Baker's, even though only Able's rock hits the bottle. If no rock had hit the bottle, then neither of them would have thrown.

We could say so, and sometimes we might: but in that context of inquiry, it is no longer acceptable to say that Able's throw caused the break. Able's throw did not cause *some* rock to hit the bottle: the quick-fire bottle-smashing competition did that. This sense of cause is the sense in which the soldier in the firing squad whose bullet first pierces the revolutionary's heart does not kill him, because it was so massively redundant in that context (which is one reason why firing squads are used). But in this discussion, we have confined ourselves to the sense of causation in which the soldier's bullet *does* kill the victim, even though it is very redundant in the circumstances. Likewise, there are contexts where we blame the quick-fire competition, rather than any particular throw,

[12] If you disagree, ask a friend first to first brush you with a rock and then to hit you with a rock. Why would you react differently to these events? You ought not, if they are sufficiently similar that they only fail to be identical because they fail to be numerically identical. So that is not the only reason they fail to be identical. Assuming we can set aside the possibility that your different reactions to these two actual events are due to some cumulative effect, a change in your dispositions, or some other extraneous factor, it seems therefore that an actual strike fails to be identical with a counterfactual brushing, at this level of description.

for the bottle breaking. This shows once again the flexibility of selection, which we have already discussed in some detail (4.4.3 and 4.6); and it also shows once again how the intuitive situation in cases of preemption is not quite as clear-cut as it is sometimes said to be.

Does The Proposal Work For The Alarm Clock Example?

Let us briefly apply the proposal to another example which we used to make trouble for Lewis's influence account. My wife presses a button, which connects a circuit in the clock, setting the alarm. When I push the button later, the circuit has already been connected and I am just moving internal parts of the clock without connecting any circuits. If the alarm had not gone off, then that connection of the circuit would not have occurred; and if the connection had not occurred, then my wife would not have pressed the button. But it is false that, if that connection had not occurred, I would not have pressed the button. Note that the connection event occurring when my wife pushes the button is different from the one that would have taken place if I had pushed the button an hour later, even if its intrinsic properties are the same. To see this, note that it is possible for both to occur — for example, if the battery were removed and replaced between our respective button-pushes, disarming the alarm. Then a connection would occur when my wife pushed the button and a connection would occur when I did. This would be impossible if the two connection events were one, since then they could not be two in any world.

6.6 Trumping

6.6.1 The Problem

We have been assuming that, in all cases of preemption between actually occurring events, there is a failure of intermediary events in the preempted causal chain. However, there might be cases where:

- c_1 and c_2 both occur;
- each would suffice to cause e in the absence of the other;
- there are no intermediary events;
- c_1 causes e and c_2 does not cause e.

Suppose, with Schaffer, that Merlin and Morgana both cast spells to turn the Prince into a frog at midnight. Suppose further that Merlin casts his spell earlier in the day, and that "it is a law of magic that the first spell cast on a given day match the enchantment that midnight" (Schaffer 2004a, 59). Hence, had Merlin preferred a toad, the Prince would have become a toad; but had Morgana preferred a toad, the Prince would still have become a frog, due to the priority of Merlin's frog spell. But in fact they both felt that the appropriate amphibian for the Prince to become was a frog. Had Merlin not cast his spell, the Prince would still have become a frog, due to Morgana's spell; and likewise, had Morgana not cast her spell, the Prince would still have become a frog, due to Merlin's spell. This clearly presents a problem for Lewis's original analysis in terms of dependence; it also creates a problem for the influence account, if we assume — as we are free to — that magic is a fairly coarse art, so Merlin can't vary how or when the Prince becomes a frog.

On the first pass the Reverse Counterfactual analysis appears to suffer equally, and for the same reasons:

> ...there is neither a failure of intermediary events along the Morgana process, ...nor any would-be difference in time or manner of the effect absent Merlin's spell, and thus nothing remains by which extant [counterfactual accounts of causation] might distinguish Merlin's spell from Morgana's in causal status.
>
> (Schaffer 2004a, 59)

The Reverse Counterfactual analysis appears unable to distinguish trumping from trumped cause. For the modified condition proposed in 6.5.1 above only works if there is a failure of intermediary events along the preempted process, but not in the preempting process. And by stipulation, there is no such failure in a case of trumping.

I think the best way to handle trumping is to ask for further justification of the claim that the trumper, and not the trumped event, is the cause of the effect. Stipulation is not enough. We need to know exactly why this fails to be a case of symmetric overdetermination: why we should acquiese to the assertion that Merlin's spell caused the transformation, and not Morgana's. In Schaffer's example, there is a law of magic which ensures that the spell and the enchantment match; but simple matching is a weaker notion than our intuitive notion of causation.

I think there are two ways we might justify the claim that Merlin's spell is effective and Morgana's ineffectual. The first is that there is some sort of

underlying mechanism by which Merlin's spell takes priority (and over which the law generalises). The second is by appeal to some notion of influence, close to Lewis's: some contrast exists between the way possible differences in Merlin's spell match up to possible differences in the effect, and the way possible differences in Morgana's spell do. I shall argue that neither option preserves the trumping objection as it applies to the Reverse Counterfactual. Finally I shall suggest that, if the trumping objector takes neither of these two options, trumping collapses into symmetric overdetermination.

6.6.2 Supressed Mechanisms

What justifies the claim that Merlin's spell, but not Morgana's, causes the Prince to turn into a frog? One thought is that there is some mechanism — in the loosest sense of the word — by which a spell operates on its victim, and this mechanism operates for Merlin's spell but not Morgana's. We could think of this in any number of ways. Perhaps the fact of Merlin's earlier utterance renders Morgana's spell entirely impotent, an empty utterance — a spell only in the sense that it is the right form of words to have had magical consequences in other situations. Then Morgana's spell becomes like my button-push in the alarm-clock case; somehow it does not do whatever it would have to do in order to have an effect. Or perhaps Morgana's spell is potent, but as it issues forth towards the Prince, it is blocked, prevented from reaching him, either by Merlin's spell or by something triggered by Merlin's spell.

In other words, we might say that Merlin's spell causes the Prince to turn into a frog in virtue of some *suppressed mechanism*, suppressed from our description but underlying both the law of magic and our causal claim. Such a response would give us a clear intuitive reason to say that Merlin's spell, but not Morgana's, acted on the Prince. But it is pretty clearly in tension with the claim that there are no intermediary events which a theory could deploy in defence against the counterexample. And if it is the difference between cause and trumped event, it is clear that the Reverse Counterfactual account can handle trumping just as it handles preemption. For then there is at least one event between Merlin's casting and the Prince's turning, though we don't know what it is: we can gloss it as the spell somehow taking hold of the Prince. If the Prince hadn't turned into a frog, this event would not have occurred. And if this event — Merlin's spell taking hold of the Prince — hadn't occurred, then Merlin wouldn't have cast his spell. Morgana's spell fails the last condition, though: if Merlin's spell hadn't taken hold, there is no reason to suppose that

Morgana didn't cast.

Let us take it, then, that the trumping objector will reject the claim that there is some mechanism, suppressed in the description of the law of magic, in virtue of which Merlin's spell rather than Morgana's causes the Prince to turn into a frog.

At this point it is worth noting the magical nature of this trumping case. It is all too easy to suppress a mechanism when the example is entirely magical, or to simply stipulate that there is no mechanism. The question is not whether cases like this are logically possible: they seem to be. The real question is whether they qualify as cases of causation, and this is where I have sought to exert some pressure. If it is denied that any (or stipulated that no) mechanism mediates Merlin's casting and the Prince's transmogrification, then, I suggest, we must look elsewhere for our justification of the claim that Merlin's spell is has some sort of causal priority over Morgana's.[13]

6.6.3 Influence

If we reject the suggestion that there is some suppressed mechanism, then in what sense does the law give Merlin's spell priority? An obvious thought is that what Merlin does matters: if he prefers a toad, the Prince will become a toad. Whereas, what Morgana does, does not matter: if Morgana prefers a kangaroo, the Prince will not become a kangaroo unless Merlin also prefers a kangaroo. And there is no reason to suppose that Merlin and Morgana's wishes vary in tandem.

Clearly, though, this response violates the second of Schaffer's claims against trumping: that there is no "would-be difference in time or manner of the effect absent Merlin's spell" (Schaffer 2004a, 59). Thus it offers a foothold for counterfactual accounts to distinguish trumped event from trumper. Lewis presses this line:

> If Merlin's first spell of the day had not been prince-to-frog, but rather king-to-kangaroo, the transmogrification at midnight would have been correspondingly altered. Whereas if Morgana's trumped spell had been, say, queen-to-goanna (holding fixed Merlin's spell

[13]Stipulating that there is no suppressed mechanism is rather less plausible in real-world examples, such as Lewis's case of a Sergeant and a Major simultaneously shouting "Advance!" (Lewis 2004a, 81). The Major trumps the Sergeant. To me it seems pretty difficult to deny that there is *some* mechanism, some way, in which the Major's order takes priority over the Sergeant's, although Lewis claims that the denial is epistemically possible.

> and the absence of any still earlier spell) what happened at midnight would have been exactly the same as it actually was: The prince would have turned into a frog, and that would have been all.
>
> (Lewis 2004a, 93)

We have seen that influence is not generally either necessary or sufficient for causation. In cases of trumping, however, it does seem to be one way in which we might distinguish trumper from trumped event.

To make trouble for the Reverse Counterfactual, the proponent of the trumping objection must assert that $\sim E >\!\sim C_2$, which is equivalent to $\sim (\sim E \geq C_2)$.[14] Translated back into English, and applied to Schaffer's example, we get this claim: it is false that, if the Prince had not turned into a frog, Morgana might have cast her frog spell.

This rules out one way in which the objector might try to characterise the distinction between trumped event and trumper. The objector cannot admit that the effect could have been different, while the trumped cause might have been the same. But how plausible is that? Merlin cast first; he could have cast a toad spell; then Morgana might still have cast her frog spell, yet the Prince would not have turned into a frog. In worlds talk, among the worlds where the Prince does not turn into a frog, are some where Merlin is still angry, but thinks "Hmm... I'm bored with frogs; I think it shall be a toad today." And in some of those worlds, Morgana's spell remains a frog-spell, yet the Prince becomes a toad, as Merlin wishes. At least, that arrangement of worlds would be a reason to accept the claim that Merlin's spell has causal priority over Morgana's. If that is the reason, then the Reverse Counterfactual fails for Morgana's spell, because as we have just seen, it is false that worlds where the Prince fails to become a frog and where Morgana does not cast her spell are all closer than any where she still casts her frog spell.

The objector could seek to deny that the Reverse Counterfactual is true for Merlin's spell, but that is not plausible either: if the Prince had not turned into a frog, then surely Merlin at least would not have cast his frog spell. To use the worlds locution, worlds with earlier sorcerors or different laws seem to be further away.

It seems, therefore, that distinguishing Merlin's spell from Morgana's by their variable influence on the effect makes the Reverse Counterfactual true

[14] Removing double negation from $\sim\!\sim C_2$.

of the trumper but false of the trumped event. At least, it seems so in this example. The objector might respond, however, by tightening the example so as to preserve some sort of intuitive distinction between trumper and trumped event, while removing the resources to which the Reverse Counterfactual appeals. Let us therefore focus our attention on this effort.

Although the Reverse Counterfactual is not appealing directly to the notion of influence, it might be objected that the foregoing argument relies on trumpers being counterfactually more fragile than trumped events, in a certain way.[15] But some sorts of cause-effect pairs are quite rigid: they either happen or don't. Suppose we have a rigid cause-effect pair trumping a cause which admits of much more variable effects: then supposing the effect away would mean supposing the cause and the trumped event away, but we could characterise the causal priority of trumper over trumped event by mentioning all the ways the *trumped* event could have been different, without having any bearing on the effect.

An example will make this clearer. I want a coffee, and I am about to go out to get one, when the fire-alarm rings. A fire-alarm is pretty much a one-trick pony: all it ever does is get people to leave the building. Desires, on the other hand, can vary widely, and cause a wide range of actions. I might have wanted biscuits, which are on a tin by the window; or tea, which is available along the corridor: but I would still have left, because the fire alarm trumps my desires.[16] This is an objection to the Reverse Counterfactual because it exhibits an intuitive way to cash out the priority of the trumper over the trumped event, while ensuring that the Reverse Counterfactual is true of both, and is therefore useless to distinguish between them. If I had not left, then I would neither have heard a fire alarm nor wanted coffee. Yet the fact that I would still have left even if I had wanted biscuits or tea seems a good reason to say that the fire alarm and not my desire for coffee made me leave (and indeed to leave the coffee out of any explanation I might offer to an inquisitor waiting outside the building).

As it stands, however, this modification of the example does not establish causal priority. If we honestly restrict the possible variations in the trumper so severely, and hold one event fixed, then no matter how you vary the other,

[15] This is another variation on the Lipton objection discussed in 4.6.2.

[16] Both the alarm and the desire for coffee admit of variation in time, but let us ignore this, because it does not distinguish between them. Had I left earlier, at least one of them would have happened earlier, but we can't tell which; had I left later, both would have happened later. (The situation is the same Lewis's way round: if I had wanted coffee earlier, I would have left earlier, and if the alarm had gone off earlier I would have left earlier.)

the outcome is the same. Hold the fire alarm fixed and vary my desires, and I will still leave the building; hold my desire for coffee fixed, and vary the fire alarm the only way we are allowing — that is, suppose it didn't go off — and again, I leave the building.

The point is a little subtle, perhaps, but I think it is important. At first, it looks like there is good intuitive force behind the thought that, *whatever* my desires had been, I would still have left when I heard the alarm. Indeed there is: my desires are irrelevant when the alarm is ringing. But on its own, that does not establish the causal priority of the alarm, only the inefficacy of my desires. That priority is only established through an implicit contrast with the case where the alarm — somehow, or perhaps even *per impossibile* — ellicits a different effect from the actual one. It boils down to a contrast between two counterfactual scenarios. In both I stay inside; in one, the alarm still rings, and in the other, I still want coffee. Clearly there is a sense in which the latter is a more likely scenario, a closer world. In this sense, if I had not left the building, I might still have wanted coffee, but there would have been no alarm. For me to ignore an alarm would take a serious departure from the actual world — deafness, obsession with earplugs, a hostage situation, or some such.

So appealing to ineffectual variation in the trumped event will not provide a distinction — either for our intuitions or for Lewis's influence theory — between trumped and trumping event, without the crucial admission that $\sim E \geq C_2$ — that if the effect hadn't happened, the trumped cause still might have done, even if the sense in which it might have done might be rather artificial, restricted, determined by the context of inquiry. The artifice is due to the strange nature of the case. If you asked me why I had left the building, I would probably say, "Well, there was a fire alarm, and I wanted a coffee anyway." The case presents itself as one of overdetermination. So, like the sense in which I might have stayed inside but still have wanted coffee, the sense in which this is *not* a case of symmetric overdetermination is somewhat artificial.

Notice, finally, that Lewis's account does not appear as well placed. Variations in the fire alarm will not produce variation in the effect, but equally small variations of my desire will do so as well — in particular, if either occurs slightly earlier, I will leave the building slightly earlier. We can modify the Merlin example to have a similar property as the fire-alarm case: suppose Merlin has only one spell, being an old dog, whereas Morgana has a load of

new tricks she could use, but on this occasion chooses to use the frog spell. Lewis's account struggles: if Merlin had not cast his frog spell, but Morgana had cast hers, then the requirement that Merlin is a one-trick pony means that the Prince would still have turned into a frog. For worlds where Merlin varies his spell slightly are further than worlds where he does not cast a spell at all. By starting with the effect, the Reverse Counterfactual avoids this difficulty, and, I suggest, comes closer to capturing the intuitive distinction we draw. Merlin's spell does not "make the difference", in any ordinary sense, because of the presence of Morgana's spell. If we want to recapture the intuitive notion of difference making, in order to preserve the suggestion that Merlin's spell takes priority, then we have to contrast the case where the effect doesn't happen and Merlin still casts with the case where the effect doesn't happen and Morgana still casts. Morgana's spell is not a reliable difference between the actual and the former case: had Merlin's spell not worked, we just don't know what would have become of Morgana's. She might have cast, or she might not. But Merlin's spell is a somewhat more reliable difference between the actual and the latter case. Had the effect not occurred, Merlin would not have cast; supposing Morgana to have cast merely requires that we suppose something else to have prevented her spell from working, when in actuality Merlin's spell prevents it.

6.6.4 Trumping and Symmetric Overdetermination

Our objector might, however, push one stage further, and insist that neither the effect, nor either of the putative causes, admits of enough variation to run this sort of response. Suppose there is only one spell, and it turns people into frogs. But then the law of magic, that the first spell uttered in a day takes priority, would be replaced by a law which said that if anyone uttered a spell that day, the target would become a frog at midnight. There would be no sense in which Merlin's spell took priority over Morgana's. Again, if my heart has only ever known the desire to leave the building, then how do we make sense of the suggestion that the alarm had priority over the desire to leave the building?

Each of these cases is a case of symmetric overdetermination: two sufficient causes occur, and are indistinguishable with regard to the effect. If there is no suppressed mechanism, and no counterfactual difference, then it is hard to see what justifies the claim that one rather than the other is the cause. For it is hard to see what the claim of priority amounts to. It is neither a claim about

the way the candidates are differentially linked up to the effect, nor about how they vary differentially with the effect in counterfactual scenarios. The case displays all the characteristics of symmetric overdetermination. And as we are about to see, the Reverse Counterfactual offers a respectable account of symmetric overdetermination.

6.7 Symmetric Overdeterminaton

6.7.1 Lewis's Position

Lewis's early view of symmetric overdetermination is dismissive. Because our intuitions about symmetric cases are vague, such cases are not useful for testing a theory of causation against, and may be left as spoils to the victorious theory (Lewis 1986a, 171 — note 12). His view subsequently became more sophisticated, classifying and saying more about different sorts of symmetric cases, but his attitude to the residual cases remained largely unchanged.

His relaxed attitude arises in part from the following reasoning.

> I should dispel one worry: that if we were ever to decline to count redundant causes as genuine causes, we should be left with gaps in our causal histories — no cause at all, at the time when the redundant causes occur, for a redundantly caused event. For consider the larger event composed of the two redundant causes... Whether or not the redundant events are genuine causes, the larger event will be there to cause the effect. For without it — if it were completely absent, with neither of its parts still present, and not replaced by some barely different event — the effect would not occur.
>
> (Lewis 1986a, 212)

Whatever Lewis might be able to say about various sorts of symmetric overdetermination, his underlying attitude to residual cases is that causal histories can always be preserved from gappiness by appealing to all overdetermining causes, and saying that if *that* event had not occurred, the effect would not have occurred.

We need to be careful to distinguish two claims, one of which I will argue is intuitive, the other problematic. The intuitive claim is that, in overdetermination cases, both redundant causes cause the effect. Two bullets pierce the heart, and we want to say the same thing about both of them; as Lewis

anticipates, we don't want to say neither caused death, and risk a gap; so we might be inclined to say they both did. The less intuitive claim is that in symmetric overdetermination, two redundant causes together form one larger cause. This is the Lewisian solution, and I will seek to explain why I think it is problematic, and distinct from the intuitive view to which it bears similarity.

What is the nature of this "larger event" composed of two redundant causes? I lifted the answer from the previous quotation, and now give it:

> I mean their mereological sum. Not their disjunction — I do not know how a genuine event could be the disjunction of two events both of which actually occur. It would have to occur in any region where either disjunct occurs. Hence it would have to occur twice in one world, which a particular event cannot do.
>
> (Lewis 1986a, 212)

Why does Lewis mention disjunctions? — Because, on Lewis's analysis, the most natural first thought would be that the disjunction is causal. For to say that neither c_1 nor c_2 occurs is to say that $\sim (C_1 \vee C_2)$. Lewis can, and does, suggest that we should ignore this point. But the argument proceeds from a theory of events. It does nothing to say whether or why $\sim (C_1 \vee C_2) > \sim E$ is false; and nor should it, since it is true (assuming the overdetermining events are not themselves collectively redundant).[17] But Lewis cannot allow disjunctive events as causes, and not only for the reason he gives above. It would allow all sorts of spurious cases of causation, such as my walking or my talking causing my arriving (cf. Lewis 1986a, 190). Lewis argues that disjunctive events cannot be allowed, because then they would be wholly present whenever their disjunts were, meaning a numerically identical event could occur twice in one world. Obviously any event can be represented with a disjunction; he disallows disjunctive events whose disjuncts are "highly varied" (Lewis 1986a, 190) — that is, occupying disjoint spatial temporal regions. This leaves him with a problem regarding the most obvious application of his analysis to overdetermination. Although $\sim (C_1 \vee C_2) > \sim E$ is true, the disjunction of c_1 and c_2 is not the cause of e, because it is not an event. On the (not entirely trivial)

[17] It would not be any help to reformulate thus: $((\sim C_1 \& \sim C_2) > \sim E)$. First, perhaps pedantically, this does not have the form of the Lewisian analysis, since the antecedent of the counterfactual is not a negated proposition that an event occurs. Second, what would be the cause, on this account? Presumably the event whose occurrence is designated thus: $\sim (\sim C_1 \& \sim C_2)$. But it is not clear what is meant by saying, "The cause of e is not the non-occurrence of c_1 and the non-occurrence of c_2". Whereas the disjunctive version yields a version that is more readily intuitively grasped: "The cause is c_1 or c_2."

assumption that all events have causes which are also events, a replacement must be found.

Lewis proposes a replacement, as we have seen. He asserts that the mereological sum of the two redundant causes is the cause. The sum of $c_1$and c_2 forms a further "sum event" s, such that $\sim S >\sim E$. Presumably the thought is that conjoining events to give bigger events is not a problem in the way that disjoining them is: it does not yield the unfortunate result that the same event can occur twice in the same world. But given the difficulty of disjunctive events, it is precarious to depend on the hope that sums will not be problematic. And indeed, on reflection, they do seem to be problematic.

Consider an event c_1 causing some other event e. Find any other actual event c_2 whose absence would not itself cause e. Let s be the mereological sum of c_1 and c_2. Now $\sim S >\sim E$ is true, and since Lewis offers a sufficient condition for causation, presumably that means that s causes e. But the only restriction we placed on c_2, was that its absence not prevent e. Other than that, the extra could be anything at all: so it might be something entirely irrelevant. Lewis wants to rule out my talking *or* walking as a cause of my arrival, because he denies disjunctive events (and events are the causal relata, for Lewis). But on the mereological summing picture, my talking *and* walking qualifies as a cause my arrival, because if the entire talking-walking event had not occurred, I would not have arrived. Surely, though, walking and talking didn't get me here: the walking did that by itself, and talking played no part.

Lewis could further require that the sums be spatiotemporally continuous, but this would not help, for two reasons. First, a cause of some event e will be spatiotemporally contiguous with many other events that do not cause e. (Otherwise, either there are gaps in space and time, or else everything causes e.) If two events are contiguous, their sum will be continuous. So many non-causal sums would still be admitted as causes. Second, many overdetermining causes are not spatiotemporally contiguous. Two assassins fire at just the same time, and both cause the president's death, though the action of either alone would have sufficed. But the assassins are not holding hands: they are in quite different places, and certainly not spatiotemporally contiguous. So this further stipulation would rule out many overdetermining causes.

Realising this, we might ban *arbitrary* mereological sums from being causes, as they are banned in Lewis's own theory of reference, for example (Lewis 1984). But then, I think, we have no response to the problem of symmetric overdetermination. For the mereological summing required by that response

will surely violate any useful restriction on arbitrary sums: redundant causes can be events of very different kinds.

Let us return to the notion, which I claimed was intuitive, that in symmetric cases, the two redundant causes both cause the effect. I hope the difference between this claim and Lewis's response is now clearer. It is one thing to say that two events *both* caused a third; it is not so very different from making any other claim about two events, for example, that they both took place in Turkey in 2003. It is something else entirely to roll two events together, and claim that they form one larger event, which causes a given effect. Lewis moves from former to the latter, but they are clearly not the same. To say that I was both married and drunk on the beach in Turkey in 2003 is perfectly acceptable; but to sum these events would be a slur on my character. For there is no event of my wedding and drunkenness on the beach. We cannot sum arbitrary events to create larger events, any more than we can disjoin them; for the mereological sum of two events is not necessarily an event. This is an intuitive constraint which any theory of events should respect, at least if it is to be used to analyse other concepts subject to common judgement such as causation. The sum of my wedding and drunkenness on the beach is caused by nothing, and causes nothing — though clearly, both events had causes and consequences. Perhaps arbitrary summing is to be disallowed, then. But whatever restrictions might be required to prevent arbitrary summing being a serious problem for any event-based theory of causation, Lewis's solution to overdetermination surely violates them, as I have argued, because it does not restrict the sorts of things that can be summed.

6.7.2 The Reverse Counterfactual and Symmetric Overdetermination

In cases of symmetric overdetermination, where c_1 and c_2 have equal claim to cause e, it is the case both that $(\sim E > \sim C_1)$ and that $(\sim E > \sim C_2)$. Consider Able and Baker throwing so that their rocks strike the bottle at the exact same moment. If the bottle had not smashed, Able would not have thrown, and if the bottle had not smashed, Baker would not have thrown. Neither would have thrown. Applying the Inference test: an intact bottle would have led you to infer that neither Able nor Baker threw accurately. And in worlds talk, the nearest worlds where the bottle does not smash are worlds where neither Able nor Baker throw accurately. The Reverse Counterfactual is true of both Able's throw and Baker's.

We could apply the condition proposed in 6.5.1, but because the case is symmetric, that will not yield any difference between the two throws. And nor should it, of course. Note that the proposed condition allows both events to be causes. For each throw, there will be at least one event which is both between the throw and the break, and which the Reverse Counterfactual connects to both the throw and the break. This event is the impact of the relevant rock. Moreover, the impact itself satisfies this condition trivially, since there is no further event between the simultaneous impacts and the break.

The advantage of the Reverse Counterfactual account over Lewis's response to symmetric overdetermination is that we can neatly avoid all this trouble arising from the disjunctive antecedent of the Lewisian counterfactual, or from the mereologically suspect substitute. Once again, the reason is that we have two true counterfactuals, where Lewis has two false ones. We can simply stick to the view that if c causes e then $\sim E >\sim C$, and if there is more than one c fitting that bill, and nothing further to choose between them, then e has more than one cause. We can admit that $(\sim E >\sim (C \vee D))$. But we do not thereby need to commit to admitting problematically disjunctive events. The disjunction in that claim can be taken at face value, as a disjunctive claim about two events; we remain free to deny, with Lewis, that disjunctive events make much sense. For we are free to claim that the causes are the events represented by each of the disjuncts, and that the disjunction simply represents a claim about both of them — in the same way that disjunctions normally do, without causing problems. Of course this was not an option for Lewis: for the disjuncts themselves were not individually causal, on his account; only the disjunction was causal. Whereas on the Reverse Counterfactual account, the disjuncts are both causes. We can speak about one or the other without committing to the causal efficacy of their mereological disjunction. And their mereological sum need not be mentioned at all.

Note that on this account, there is a distinction between joint causes and overdetermining causes. Our condition on joint causation (4.5.2) required that causes feature ineliminably in the jointly causal set. Since overdetermining causes are each sufficient, they would feature eliminably in any set containing more than one of them. The two bullets which simultaneously killed the president could each be removed, and the president would still die. I think this is a positive feature of the account, since there does seem to be an intuitive distinction between symmetric overdetermination and ordinary joint causation.

The Reverse Counterfactual evidently offers a much simpler account of symmetric overdetermination than any available to Lewis, since it requires no fancy footwork concerning the metaphysics of events. I think it also reflects our intuitive judgments better. For when Able and Baker throw rocks at the bottle, we do not say that the *event* of both of them throwing caused the smash. We say something that sometimes looks similar, but really is not: that both of them throwing caused the smash. We do not thereby mean to imply that both of them throwing was one event. There might be contexts where we do something like that — for example, if there are many throwers, and we say the bottle smashed in a hail of rocks. But that is quite a different causal claim: I am not obliged to make it in order to say that Able and Baker both smashed the bottle. I might regard Able's throw as something quite distinct from Baker's, but still say, in a symmetric case of overdetermination, that they both caused the smash. Allowing the mereological sum of Able's and Baker's throws to be causal means lifting restrictions on summing and opening the door to an awful lot of intuitively non-causal sums. For Able and Baker's throws might be quite different; whatever restrictions there are, we can surely think of redundant cause pairs that would violate them. In general, redundant causes need not be similar in any particular respect. The clarity of this point is a nice feature of the case of Camper, to which we now turn.

6.7.3 The Mysterious Case of Camper

> The night before Camper is to set off into the desert where the water in his canteen will be his only available drink, Poisoner puts a fatal poison into the canteen. Early the next morning, Spiller maliciously empties the canteen, in complete ignorance of what Poisoner had done the night before. Camper then sets off into the desert, in all innocence, and dies of thirst. Who killed Camper?
>
> (Philosophy pre-interview test, King's College, Cambridge. Adapted from Mackie 1974, 44-6.)

When my seventeen-year-old self sat this test, I answered that Camper killed himself, by failing to check his water bottles as he should have. It has long been an ambition of mine to improve on this answer, and I will take this opportunity to do so. At the same time we will be able to apply some of our machinery to an intuitively tricky example.

Mackie argues, against Hart and Honore (1985), that Spiller caused the death. He reasons as follows.

> ...the event which was the traveller's death was also his death from thirst, and we must say that the puncturing of the can caused it, while the poisoning did not. ...it is the chain puncturing—lack-of-water—thirst—death that was realized, whereas the rival chain that starts with poison-in-can was not completed.
>
> (Mackie 1974, 46)

Therefore, Spiller caused Camper's death; Poisoner would have caused it if Spiller had not. According to Mackie, then, this is a case of preemption. Contrarily, I shall argue that it is a case of symmetric overdetermination.

Mackie's reasoning is objectionable because the causal chain he maintains was realized was not, in fact, realized. The chain "puncturing—lack-of-water—thirst—death " is the causal chain that *would* have occurred, had Poisoner not previously put poison in the water cans. To say, as Mackie says, that puncturing caused the lack of water, and the lack of water caused the death, is an equivocation on the term "water". The water which Spiller poured away was fatally poisonous. To say that depriving Camper of this fatal liquid caused him to die is implausible; it only gains plausibility if we call that liquid "water", for everyone knows that without water, Camper will die of thirst. We could bring the point out by denying that the fatal liquid was water, or denying that it was pure or good water, or some such; but I do not think we need to rely on any such claim to make the central point.[18] That point is that the realized chain from Spiller's action goes puncturing—lack-of-poisonous-water—death. This makes rather less sense, because lack of poisonous water would usually be thought to be a condition for life, not a cause of death. Yet lack of poisonous water is all that Spiller's actions caused (assuming Poisoner had been thorough, and poisoned all the water cans).

So what should we say about the case of Camper? A number of considerations spring to mind. First, the *time and manner* of Camper's death is no different than it would be if only Spiller had acted. If we hold Spiller's action constant, then Poisoner's actions had no effect on the time and manner of the death. On the other hand, if we hold Poisoner's actions constant, then

[18]I understand that, in interviews following the test at King's, applicants are sometimes asked how their answer might change if Poisoner's place was taken by Petroler, who empties the canteens and refills them with petrol, which is toxic. I am in effect suggesting that our answer should not change, whether Poisoner or Petroler acts that night.

Spiller's action did not make any difference to *whether* the death took place. The same point applies in reverse, of course: if we hold Spiller's actions constant then Poisoner's earlier activity is irrelevant to whether the death occurs as well as to how it occurs. But here, I think temporal order is relevant to our intuitive judgements. At the time at which Poisoner acted, Camper's fate was not sealed, because nobody had fatally sabotaged his water supply. From the point after Poisoner acted, however, it was. Whereas when Spiller's turn came, he might as well not have bothered; indeed maybe he would not have bothered, had he known that Poisoner had already acted. Poisoner might also have desisted if he had known of Spiller's future actions, of course; but in doing so he would be taking a risk, with regard to his murderous intent. On the other hand, the way things actually turned out, no poison ever entered Camper's bloodstream. Poisoner might be an expert in his field, who prides himself on picking just the right poison for the job; but in this case he would have no reason to boast.

That is a barrage of intuitions. We could produce variations on this example which removed the niceties of this particular case, in order to push one intuition or another. But these niceties are interesting because the case is so realistic: it could happen just like that. We can imagine the lawyer for Spiller arguing that Spiller's action made no difference to the fact of death, only to the manner; and we can imagine Poisoner's defence arguing that his poison was rendered irrelevant by Spiller's later action, and pointing out that no gram of the fatal substance entered Camper's bloodstream. The prosecution would presumably wish to secure a conviction against both Spiller and Poisoner, but the claim that they both caused Camper's death will be disputed by both defence teams (since each will dispute the claim that their client caused the death at all). How are the prosecution to secure a conviction?

It seems to be emerging that influence is of little help here. It may play a role in Poisoner's defence: the manner of death was influenced by Spiller's actions, but not by Poisoner's. But denying that influence and causation are the same thing will be a key part of Spiller's defence: for that case would turn on the fact that Spiller's actions merely changed the manner, and not the by-then-inevitable fact, of death. The defence would be arguing that Spiller's actions were like the pudding discussed previously: they delayed the death (if, as seems plausible, it takes longer to die of thirst than poison), and changed its manner; but it was going to happen anyway. If the prosecution accept that causation is influence, Poisoner goes free; if they deny it, they must find a

substitute analysis, or Spiller goes free.

Let us see how the Reverse Counterfactual would establish causation in the Camper case. If Camper hadn't died, then Poisoner wouldn't have poisoned; and if Camper hadn't died, Spiller wouldn't have spilled. In the possible worlds idiom, the closest worlds where Camper lives are worlds where neither Poisoner nor Spiller act. Applying the Inference Test, if Camper had made an uneventful trip across the desert, we would infer that his water supply had been both sufficient and wholesome. The prosecution now have a basis from which to argue that both murderers caused the death, because they can establish that each event satisfies a necessary condition for being a cause of the death. Of course, a necessary condition will not prove causation: for practical as well as theoretical purposes, we need a sufficiency component to our analysis. That shall be the topic of the next chapter. But given the discriminatory power of this necessary condition, establishing that an alleged cause meets it could be important and useful.

6.8 Summary

We have explored the consequences, for various overdetermination problems, of reversing the counterfactual standardly used in counterfactual analyses of causation. This trick has yielded an account which, without any further modification, can deal with cases where a cause has a non-occurring back-up. Preemption in which both candidate causes occur can be handled with a relatively minor addition to the basic account, an addition which has sound intuitive motivation. Trumping requires no further amendment. In cases of symmetric causal overdetermination, the Reverse Counterfactual analysis concludes that all the overdeterminants are causes, which I have argued is a neater account than Lewis's. Finally we imagined ourselves in court following Camper's murder, and sought to apply the Reverse Counterfactual approach to a hard case of overdetermination, in which both the *sine qua non* condition and the notion of influence would have been of little help.

Before we move on, let me reiterate a general point with which we began the chapter. In cases of redundancy, it seems more profitable to think of causes as counterfactually sufficient for their effects than to think of them as counterfactually necessary. Perhaps the idea that causes might be counterfactually necessary has arisen from confusing the plausible suggestion that *some* cause is necessary for a particular effect with the distinct suggestion that a *partic-*

ular cause is necessary. The Reverse Counterfactual makes particular causes counterfactually sufficient, given a certain background, whereas the ordinary account makes them counterfactually necessary. But prima facie, particular causes are not generally necessary for their effects: there are often other ways to bring a given cause about. Hence the falsity of the Lewisian counterfactual in cases of preemption, which I have argued the Reverse Counterfactual is well-placed to solve. To say that c necessitates e, or that c makes e happen, is not to say that c is necessary for e, but that c suffices for e: which I suggest we understand as the claim that, if e hadn't happened, then c wouldn't have happened.

Chapter 7

Conclusion

7.0 Abstract

A synopsis of the defence of the Reverse Counterfactual as a necessary condition for causation is presented. I then argue that there can be no counterfactual sufficient condition for causation. I argue that simultaneous causation occurs. Since the asymmetry of counterfactual dependence is temporal, counterfactual dependence can never offer a full characterisation of causation, no matter how sophisticated a thesis might be proposed of the asymmetry of counterfactual dependence. Thus there is more to causation than counterfactual dependence among particular events. I finish by compiling the more complicated versions of the Reverse Counterfactual which are developed in earlier chapters, along with other important claims which have been defended, in a concise summary.

7.1 Synopsis

The substantive discussion began in Chapter 2 with an outline of Lewis's semantics for counterfactuals. I sought to understand the relation between three theses: the asymmetry of overdetermination, the asymmetry of miracles and the asymmetry of counterfactual dependence. I argued that they supported each other in that order, and we considered various internal criticisms: that Lewis's semantics requires the truth of some backtrackers; that the asymmetry of miracles does not hold, thanks to the possibility of "Bennett-worlds", convergence to which is as easy as divergence from; and that the asymmetry of overdetermination fails when we consider the implications of statistical mechanics, as illustrated by "Elga-worlds". In Chapter 3, we considered what independent case exists (aside from Lewis's semantic theory) for an asymme-

try of counterfactual dependence. It is commonly thought that backtrackers and foretrackers are true under different resolutions of vagueness. But I criticised two common reasons for thinking so. Backtrackers and foretrackers are compatible, and may be mixed without any peculiarly pathological logical consequences. And the grammatical awkwardness associated with backtracking expressions does not show that they are false. On the other hand, independent reasons exist to deny that counterfactual dependence is asymmetric — at least, to deny that it is so strongly asymmetric as to yield the result that almost all backtrackers may be written off as false in normal circumstances. I proposed a method to assess counterfactuals by deploying our ordinary capacity to make and assess inductive inferences (having argued that other methods are inadequate). The Inference Test claims that our acceptance of $A > C$ and of the inference from A to C ought to covary, where that inference accords with our actual standards, but concerns the counterfactual situation in which A would occur. This test is circular as an analysis, but it was defended as a simple, intuitive test which deploys skills in which we are practised and display considerable reliability — our skills to make and assess inferences.

Thus prepared, we turned to the main topic: causation. Lewis's counterfactual is widely accepted as at least a sufficient condition for causation, making causation radically unselective: the presence of oxygen is the cause of the flame as much as the strike of the match. In Chapter 4, the unselective orthodoxy was criticised. I argued that it does not provide either a descriptive explanation of the principles governing our selective practices, nor an explanation of why we should make such heavy use of causation to select. We examined various efforts to tack a theory of selection onto an unselective theory of causation, focusing on the effort to assimilate causal selection to selection of the explanatory cause in contrastive explanation. All these efforts were found wanting, however. The Reverse Counterfactual was introduced as a natural way to explicate the notion of *the* cause as making *the* difference to its effect. It was defended as a necessary condition which causes but not mere conditions meet. The Reverse Counterfactual was promoted over competitors for several reasons. It is considerably simpler. It achieves considerable theoretical unification, depending on context only in the way that all counterfactuals do, and not in any hard-to-explain secondary way. The context-sensitivity, or flexibility, of selection is thus assimilated to the context-sensitivity of counterfactuals. It may be hard to explain exactly how counterfactuals themselves are sensitive to context, but that is a problem we already have, in giving a coun-

terfactual account of anything. Finally, the approach I have advocated makes selection central to causation, offering an account of why we use causation to select in such explanatory, predictive, manipulative, moral and legal contexts. Any account which purports to be an analysis of our ordinary concept, and yet which does not make selection central to causation, must shoulder a heavy explanatory burden with respect to the frequency and gravity with which we apply causal concepts to select.

Chapter 5 extended the argument to cover selection of the cause in preference to other events in the same causal chain. I argued that causation is not transitive. Reasons for the widely-held view that it is transitive were considered and rejected. Two circumstances in which causation seems to fail to be transitive were distinguished: those where the cause fails to be sufficiently proximate to the effect, and those involving double-prevention. I argued that the Reverse Counterfactual readily distinguishes causes that are sufficiently proximate from those that are not, being true of the former but not of the latter. Double-prevention counterexamples to transitivity were considered in some detail, and it was found in each case that the first event in the sequence indeed failed to cause the last. Three diagnoses were considered, and the Reverse Counterfactual was found to have links to all of them. Finally I proposed a diagnosis of my own, arguing that in cases of double-prevention, the first event does not cause the second. So these cases are not really counterexamples to transitivity. A suggestion was made as to why causation does appear to be transitive over shortish chains of certain kinds, to the effect that a valid substitute for counterfactual transitivity is available in those circumstances which entails that the Reverse Counterfactual is true.

Lewis's account makes causes counterfactually necessary for their effects, but causes are sometimes redundant, meaning that they are sometimes counterfactually unnecessary for their effects. In Chapter 6, I argued that the Reverse Counterfactual should be read as making causes sufficient, in a certain sense, for their effects, and therefore that redundancy would not threaten the Reverse Counterfactual with falsity, as it threatens Lewis's accounts. Various kinds of redundancy were considered, to see whether the Reverse Counterfactual could further distinguish redundant causes from non-causal events. I argued that the Reverse Counterfactual can readily handle any case of preemption where the preempted event does not actually occur. Where both potential causes occur, I argued that the same strategy could be applied to events further down the respective causal chains. The chain from a preemptor to an

effect goes all the way, whereas the chain from a preempted event does not. Where it fails, there will be events which do not actually occur but which, if they had occurred, would have completed the chain to the effect. I suggested we appeal to the contrast between these events and the occurring events in the complete causal chain from cause to effect to distinguish preemptor from preempted event. This enabled me to propose a solution to both early and late preemption. Cases of trumping, however, are designed to have no intermediary events which might be used for this sort of solution. It was argued that, to set up a convincing case of trumping, either an implicit appeal must be made to some intermediary events between cause and effect, or else an implicit appeal must be made to the claim that, if the effect hadn't happened, the trumped event might have happened anyway. The Reverse Counterfactual handles either sort of case without further modification. If neither claim is accepted, I argued that trumping cases collapse into cases of symmetric overdetermination. We considered the latter, and found that the Reverse Counterfactual offers an account which is simpler, more intuitive, potentially more useful and less mereologically committing than Lewis's.

I have finished advocating the central claim of this work, that the Reverse Counterfactual is necessary for causation. In what is left of this concluding chapter we shall consider the other big question: whether counterfactual dependence, of any kind, is sufficient for causation.

7.2 Causal Asymmetries

7.2.1 The Need For A Sufficient Condition

The Reverse Counterfactual has been defended as a necessary condition for causation, and various claims have been made for its power, and the power of more complex derivatives, to discriminate between causes and non-causes. But to discriminate between causes and non-causes more generally, we need a sufficient condition as well as a necessary condition. The Reverse Counterfactual and the various derivatives we have considered clearly fail to amount to a sufficient condition for causation. The reason is obvious. The Reverse Counterfactual says that, when c causes e, it will be the case that $\sim E > \sim C$. In many cases, Lewis's counterfactual will also be true: $\sim C > \sim E$. If the Reverse Counterfactual sufficed for causation, then not only would c qualify as causing e, but e would also qualify as causing c. This is implausible: effects rarely, if ever, cause their causes. Relatedly, when c causes two effects, e_1 and

e_2, it might well be the case that at least one of the effects counterfactually depends upon the other — for example, it could be that $\sim E_1 > \sim E_2$. It might also be the case that $\sim E_2 > \sim E_1$. If the Reverse Counterfactual sufficed for causation, then on some occasions, effects of a common cause would cause each other. But effects of a common cause rarely, if ever, cause each other.

In short, the Reverse Counterfactual cannot provide a sufficient condition for causation until it has been supplemented by some means of identifying the causal order. The Reverse Counterfactual is "Reverse" with respect to Lewis's, running from effect to cause rather than from cause to effect. But unless I have a way of distinguishing cause from effect, my theory will be unable to distinguish the Reverse Counterfactual from other counterfactuals, and thus I will be unable to recommend it as a sufficient condition for causation.

The causal order matters because causation is asymmetric. In fact, there are many asymmetries associated with causation. Hausman provides a non-exhaustive list (Hausman 1998, 1), and his list could no doubt be extended by a study of a diverse bundle of literature on time, quantum physics, laws and explanation. I propose to focus on just two causal asymmetries. The first is the *metaphysical asymmetry of causation*. This is the fact that causation is an asymmetric relation.[1] The intuited asymmetry of causation seems to be of a strong rather than a weak sort. We can say that a relation R is *symmetric* iff $(\forall x)(\forall y)(Rxy \supset Ryx)$, and *weakly asymmetric* iff it is not symmetric. R is *strongly asymmetric* iff $(\forall x)(\forall y)(Rxy \supset \sim Ryx)$. Strong asymmetry implies weak, but not vice versa. Causation appears to be strongly asymmetric, in common cases: we move readily from "c caused e" to "e does not cause c", which is licensed by strong asymmetry, but not weak. I call this a *metaphysical* asymmetry to reflect the fact that it is not a matter of logical impossibility that an effect does not cause its cause, but rather a matter of metaphysical impossibility.[2]

The second asymmetry we shall examine is the *temporal asymmetry of causation*, which in its simplest form consists in the fact that effects do not precede their causes, in our common experience. In short, effects neither cause nor precede their causes.

It is possible that neither asymmetry in fact holds. It may be discovered that effects sometimes precede their causes, or that effects sometimes cause their causes. My characterisation of these asymmetries is not intended as a

[1] If it is a relation at all, of course. If not, it had better be something which admits of asymmetry in a similar way.

[2] In this formulation I am grateful to Dan Heard.

statement of unalterable conceptual necessity. Rather, it is a statement of the principles which seem to govern our concepts as we put them to ordinary use. Ordinarily, we do not see how effects could cause their causes. We might be corrected by abstruse cases from the borders of physics, but until we are, I suggest, we take something like the metaphysical asymmetry to hold. Someone alleging a case where an effect caused its cause would have some serious explaining to do, and perhaps also some extension and revision of our ordinary concept. The temporal asymmetry is even more obviously prone to exception or revision; we can imagine backwards causation, but we do not normally think it happens. Again, this shows clearly in our causal reasoning: we do not expect our current actions to have past consequences, nor the events we witness to have future causal origins. Some physicists may postulate backwards causation (cf. Dowe 1996), but again, abstruse cases merely highlight the fact that we ordinarily take an asymmetry to exist. Evidently, we use both the temporal and the logical asymmetries of causation in ordinary causal thinking. For this reason, a theory of causation should address itself to the nature of at least these two causal asymmetries, and perhaps to others as well.

Let us examine Lewis's effort to relate these two causal asymmetries, and to provide a characterisation of them in terms of counterfactuals (7.2.2). I shall argue (7.2.3) that the occurrence of simultaneous causation means that the temporal asymmetry of causation is not always present: so it cannot be analysed with the temporal asymmetry of counterfactual dependence (even if the latter were accepted). We shall consider (7.2.4) an argument from Hausman which seeks to reclaim a weaker asymmetry than Lewis proposes. Finally (7.2.5) I shall argue that none of this is really Lewis's fault: any counterfactual account will suffer similar problems, because the only obvious way in which counterfactuals might potentially display asymmetry is temporally, and causation does not always display temporal asymmetry. Thus counterfactuals cannot provide the resources to model the metaphysical asymmetry of causation — *even if* Lewis's, or any other, thesis of the asymmetry of counterfactual dependence were true. Therefore, I shall argue, no counterfactual suffices for causation.

7.2.2 Lewis's Hope

Lewis identifies two problems associated with causal asymmetry: the problem of *effects* and the problem of *epiphenomena*. The problem of effects is the problem, for an analysis of causation, of characterising the difference between

cause and effect: that is, characterising the basic asymmetry of causation. The problem of epiphenomena is the problem of telling cause-effect pairs apart from effects of a common cause. The two are different versions of the same theoretical challenge: that of distinguishing instances of the causal relation from other relations satisfying whatever theory is proposed, but failing to be cases of causation. Lewis sees this clearly, and offers a single, elegant solution: deny backtrackers. This solves the problem of effects as it arises for his counterfactual analysis, because causes precede their effects but not vice versa. And it solves the problem of epiphenomena. Suppose c causes e_1 and e_2 — the falling air pressure causes the barometer to fall, and soon it starts to rain. We might be tempted to think that if the barometer hadn't fallen, it would not have rained. But our reasoning backtracks. The reason we think that, if the barometer hadn't fallen, then it would not have rained, is that we suppose the barometer's failure to fall would have been due to a corresponding failure of the air pressure to fall. We think that, if the barometer hadn't fallen, then nor would the air pressure. But that is a backtracker, so we should reject it as false: if the barometer hadn't fallen at t, then the air pressure would already have started to fall by then. We cannot backtrack and suppose it would not have fallen. And we had no other basis for supposing the rain to counterfactually depend on the barometer's movements; so we should deny that it does. In fact we should go further, and assert the stronger claim: if the barometer hadn't fallen, it would have rained all the same. For supposing something else to interfere after the fall in air pressure to prevent the rain would be a gratuitous departure from actuality.

Lewis's ambition is to identify the direction of time with the direction of counterfactual dependence, and simultaneously to give the direction of causation and explain why causes precede their effects. The reason he gives is that causation is just counterfactual dependence, in the end, and counterfactual dependence is asymmetric. It is important that the dimension in which counterfactual dependence is asymmetric is temporal. Although Lewis's picture doesn't presuppose the *direction* of time, the asymmetry makes no sense without a prior *order* in which the asymmetry is expressed. Lewis's picture presupposes a temporal ordering. Without it, counterfactual dependence might still be asymmetric, in the sense that if $A > C$ then $\sim (C > A)$. But Lewis's arguments for that claim (from the asymmetries of overdetermination and miracles) clearly depend on events being temporally ordered. Without a temporal order, Lewis's asymmetry thesis might still be true, but the reasons for thinking so

would be absent.

If that is correct, then it appears that Lewis's solution to the problems of effects and epiphenomena rely on a questionable assumption: that causes always precede their effects. To show this, and at the same time to show why it is problematic, I shall argue that Lewis's solution does not work in cases where cause and effect are simultaneous.

7.2.3 Simultaneous Causation

I reach for the door-handle, and push it down. It turns. The mechanism is smooth and tightly-fitting. On the other side the handle I am not touching turns too. As far as I can tell, the handles turn simultaneously. Let us accept, for now, that the handles do turn simultaneously. The turning of this handle causes the turning of that one, and not vice versa. But it does not precede the turning of that one.

How, then, should we apply Lewis's solution to the problem of effects? c is the handle on this side turning, e is the handle on that side turning. They are distinct events: the handles are separate objects, not touching each other, but connected by a third object — a metal spindle running between them.[3] Lewis asserts that $\sim C > \sim E$ — that if the handle on this side hadn't turned, then the handle on that side wouldn't have. And he wants to deny that $\sim E > \sim C$ — that if the handle on that side hadn't turned, the handle on this side wouldn't have. He would do so on the basis that the latter is a backtracker. But c and e are simultaneous, so neither counterfactual is a backtracker. Lewis's solution to the problem of effects does not apply, then, to cases of simultaneous causation.

Likewise, if the effects of a common cause are simultaneous with the cause, then there will be no basis on which to deny that the two effects depend on each other, one way or the other or both ways. For the basis on which Lewis maintained the counterfactual independence of effects of a common cause was

[3]Is it doubtful whether this handle turning and that handle turning are distinct events? Granted, there is a level of description at which they are parts of the same event. However it seems fairly clear that there is also a level at which they are distinct. They come apart in different possible worlds: if there were no spindle between the handles, then when this handle turned the other would not. Yet it seems we can suppose that this handle turns in a spindle-less world, just as in our world. So the event of this handle turning has a counterpart in a spindle-less world where the other handle doesn't turn. Hence the two handles turning together, though it may be *an* event, is not the same event as just one handle turning. For another discussion of the possibility of simultaneous causation, see (Taylor 1974, 35–39): he gives the example of a locomotive pulling a caboose, which would serve my argument equally well.

that their dependence could only be asserted by reasoning back from one effect to the cause and out to the other. This was supposed to be bad reasoning because of the problems associated with backtracking, but if effects and cause are simultaneous, then backtracking is not involved.

It might be retorted that I have confused the order: it is not that reasoning is bad, and counterfactuals false, *because* of backtracking; rather, certain counterfactuals are false, and it so happens they are backtrackers. In a case of simultaneous causation, there are no backtrackers and no foretrackers, but asymmetries in counterfactual dependence persist: the effect depends on the cause, but not vice versa. — Perhaps: but what reason do we have to think so? Lewis's arguments for the asymmetry of counterfactual dependence appear to depend on events being temporally ordered, even though they do not depend on the direction of time. It is very hard (for me, anyway) to comprehend the motivation for accepting, say, the asymmetry of overdetermination as a thesis about *simultaneous* events. — Then, comes the retort, perhaps the temporal order is indeed *generally* important: the reason for the asymmetry of counterfactual dependence in simultaneous cases might depend somehow on the more general asymmetry. If causes counterfactually depended on their effects in simultaneous cases, then those counterfactuals somehow wouldn't fit into the general pattern. — Maybe there is an argument to that effect, but it does not present itself to me.

Perhaps it will be contested whether simultaneous causation ever occurs. Perhaps the handle on my side turns slightly before the other one: the metal spindle twists slightly, maybe only microscopically, and there is a slight time-delay. Two points deserve making. First, retreat to scientific sophistication is a departure from the analysis of our ordinary concept of causation. Ordinarily, we consider the events to be simultaneous; our ordinary concept of causation can treat simultaneous events as asymmetrically causally related. Even if it turns out as the result of physical theory that causes always slightly precede their effects, our concept of causation does not require this: it is quite compatible with the world turning out to contain genuine simultaneous causation.[4]

Second, retreat to scientific sophistication might compound the trouble rather than solve it. In a Newtonian framework, some causal relations are simultaneous. Consider a force, exerted on a free body, which accelerates.

[4]Epistemically compatible, I mean: of course it is possible that a world differing from ours in some fundamental physical way might be so different as not to contain anything recognisable as causation. My point is rather that we don't know that, merely from considering our concept.

The acceleration a is related to the force F and the mass m of the accelerating object thus: $a = F/m$. This is *instantaneous* acceleration. The equation does not say that force at t_1 leads to acceleration at some slightly later time t_2. Causation does not feature in the equation: if there is a causal relationship, we decide what it is independently. What should we say about this particular case? Intuitively, we judge that the force causes the acceleration, or perhaps that the force and the mass cause it together.[5] But if we accept that the effect is determined by the equation, then we accept that the causation is simultaneous. We could deny the equation. I take it we will not do that. Or we could deny that causation occurs. Russell did that, but not for this reason (cf. Russell 1917). If we take this route, then we adopt an error theory about a huge tranche of common causal judgements: any involving force. I take it that this is not an attractive option for someone seeking a conceptual ansalysis of causation.

It is therefore far from clear whether we can appeal to science to underpin an assertion that causes precede their effects. Newtonian mechanics suggests the opposite. In particular, forces, unlike the spindles connecting door-handles, do not flex or stretch. At a microphysical level, much direct causation will be simultaneous on a Newtonian view. Whether this argument can be extended to a more modern physics, I do not know. Even if it cannot, it illustrates the first point, that our concept of causation is perfectly compatible with simultaneous causes and effects. It is compatible, especially, in the sense that the metaphysical asymmetry of causation can exist when the temporal asymmetry does not.

These arguments apply equally to the problem of epiphenomena, when epiphenomena are simultaneous with their common cause. Lewis's solution to the problems of effects and of epiphenomena are both troubled by simultaneous causation.

The alleged failure of Lewis's solution to the problem of effects and, by extension, epiphenomena arises in cases of simultaneous causation. Why? If Lewis's solutions have the structure previously exhibited, then an explanation is ready. The metaphysical asymmetry of causation and the temporal asymmetry of causation are distinct. Causation is temporally asymmetric if effects do not precede their causes; that is compatible with their being simultaneous. To assert that causes always precede their effects is a stronger claim. It is this stronger claim, in conjunction with the asymmetry of counterfactual de-

[5]It doesn't matter what we take to cause what: the same point holds.

pendence, which entails the metaphysical asymmetry of causation. But the stronger claim is implausible, I have argued. Common sense admits simultaneous causation (in a way that it does not admit backwards causation), and science is not guaranteed to help.

7.2.4 Hausman's Limited Asymmetry

Hausman appears to defend a counterfactual asymmetry between cause and effect, not by appeal to time, but by appeal to inference. Hausman thinks counterfactuals are subject to a Prediction Condition, repeated here:

> **P** (*Prediction condition*) The knowledge that b would occur if a were to occur and that an event of kind **a** occurs taken by itself justifies the prediction that an event of kind **b** will occur.
>
> (Hausman 1998, 120)

Hausman uses the Prediction Condition to defend an asymmetry of counterfactual dependence which is weaker than Lewis's. Hausman argues that $\sim C >\!\sim E$ satisfies the Prediction Condition, but that $\sim E >\!\sim C$ does not. In general, knowledge of the absence of a cause allows us to predict the absence of an effect, but knowledge of absence of an effect does not allow us to predict the absence of any cause in particular. It allows us to predict that *some* cause is absent, but not which one. In other words, according to the Prediction Condition, $\sim (\sim E >\!\sim C)$. This is equivalent to asserting that $\sim E \geq C$ — that if the effect hadn't occurred, the cause might have occurred. But it falls short of asserting that $\sim E > C$ — that if the effect hadn't occurred, the cause would have occurred. Hausman takes himself to be helpful to Lewis, who sometimes appears to commit himself to the stronger claim that $\sim E > C$: Hausman helpfully points out that a weaker claim will still establish an asymmetry.

Two problems afflict Hausman's proposal. First, Lewis has a reason to prefer the stronger claim, as we saw in our discussion of preemption in 6.3.1. When c_1 preempts c_2 with respect to e, Lewis's original solution relies on an intermediary, d. It must be the case that, if d hadn't happened, e wouldn't have happened. To defend this claim, Lewis argues that, if d hadn't happened, then c_1 still *would* have happened, not that it merely might have happened: $\sim D > C_1$. We saw that the weaker claim, $\sim D \geq C_1$, would not do, since it would allow that the preempted cause c_2 and thus the effect e might still have happened anyway, undermining the claim that $\sim D >\!\sim E$ (that the effect depends on the intermediary).

Second, I have argued at length for the negation of Hausman's central premise. I have argued repeatedly that we *do* predict the absence of causes from the absence of effects. Hausman asserts, without clear argumentation, that in the absence of an effect, we are equally entitled to predict the absence of any of the causes. If the match had not lit, Hausman seems to think that we are equally entitled to predict that the oxygen or the match strike would have been absent. I have argued that the opposite is true. You see me holding a lighted match; if the match in my fingers had not been lit, you might have thought I had mis-struck it, which shows that the manner of the striking is part of the cause; but you certainly would not have thought the oxygen was absent: kitchens are not equipped with oxygen masks.

Both these arguments have already been put. I hope, then, it is clear why I think that Hausman's limited asymmetry is neither useful for Lewis nor decisive against the Reverse Counterfactual analysis.

7.2.5 The Lack of a Counterfactual Sufficient Condition

The argument against Lewis generalises. There can be no counterfactual sufficient condition for causation, because counterfactuals cannot be used to analyse the metaphysical asymmetry of causation.

There is no hope of using the temporal asymmetry of counterfactual dependence — no matter how strong or sophisticated an asymmetry can be established — to analyse all the asymmetries of causation. It may be possible to explain why causes do not follow their effects. But if we admit simultaneous causation, we cannot appeal to any temporal asymmetry of counterfactual dependence to explain any causal asymmetries persisting in simultaneous cases. As we have seen, this means that we cannot use the temporal asymmetry of counterfactual dependence to explain the *metaphysical* asymmetry of causation: for the latter asymmetry persists in the simultaneous case, while the former obviously does not. And this means we cannot use the asymmetry of counterfactual dependence — even if it were accepted — to explain the metaphysical asymmetry of causation at all.[6]

The Reverse Counterfactual analysis of causation, as it stands, is incom-

[6]Unless, that is, we are willing to countenance a disjunctive picture, on which the metaphysical asymmetry of causation when causes precede their effects arises from their order in time, while the metaphysical asymmetry of simultaneous causation arises from something else. To adopt such a picture would be to deviate substantially from our common concept of causation — which admits no such disjunction — and therefore to deviate from the project of conceptual analysis.

plete. I have proposed and defended a counterfactual necessary condition for causation. I have not proposed a sufficient condition. Now, we have a *prima facie* reason to deny that a sufficient condition can be found, at least sufficient for establishing causation in the singular case (considering just a cause-effect pair). When c causes e, it is sometimes the case that $\sim C >\sim E$; and I have argued it is always the case that $\sim E >\sim C$. If counterfactual dependence of any sort suffices for causation, then in many cases, effects cause their causes.

To provide a sufficient condition for causation, it would be necessary to account for the metaphysical asymmetry of causation; and, I have argued, this cannot be done with counterfactuals, because the only asymmetry which counterfactuals might capture is temporal. Causation is temporally asymmetric, but weakly so: simultaneous causation is possible, where the cause and effect are temporally "symmetric". And in such cases, the metaphysical asymmetry of causation persists. The lack of a sufficient condition for causation would be a serious drawback of the Reverse Counterfactual analysis, if an alternative analysis existed on which the asymmetry of causation could be analysed, and thus a counterfactual sufficient condition for causation could be given. But there can be no such account.

All this implies that there is more to some c causing some e than counterfactual dependence between c and e. But it does not follow that there is more to causation than patterns of counterfactual dependence between events generally. The Reverse Counterfactual is true much more rarely than Lewis's. Perhaps this can give us a general direction for causation. Which of a given pair of events causes the other might then be settled by appeal to this general direction, and the way the events in question fit into it. In that case, there would be something more to causation than the counterfactuals between those two events, but that something more might be the pattern of counterfactual dependence between events more generally.

Yet even if it turns out that there is more to causation than counterfactual dependence, even counting general patterns of dependence not intrinsic to causal pairs, this need not deprive counterfactual analysis of all interest. We need not despairingly conclude that causation is something else entirely, which merely has counterfactual entailments. Knowledge is sometimes thought to have a counterfactual component among components of other sorts: satisfying a counterfactual is a necessary, but not a sufficient, condition to know, on such an analysis (cf. Nozick 1981). Causation might be like that: a hybrid of

counterfactual and other components.[7] Given its complexity and the diversity of applications to which it is suited and put, that ought not be a surprise. And just as specifying the counterfactual component of knowledge is worthwhile, so is specifying the counterfactual component of causation.

7.3 Conclusion

Let us put together a concise statement of the central claim, which has been argued for concerning causation. We have the Reverse Counterfactual itself, a complication for joint causation, and a complication to handle certain kinds of redundancy. Combining these yields the following.

> **The Full Reverse Counterfactual Necessary Condition on Causation**
>
> If $c_1, ..., c_n$ cause e, then:
>
> **(i)** $c_1, ..., c_n$ include at least one actually occurring event, distinct from e [existence and distinctness requirements];
>
> **(ii)** $\sim E >\sim (C_1 \& ... C_n)$ [Reverse Counterfactual];
>
> **(iii)** $c_1, ..., c_n$ figure ineliminably, such that for every non-empty proper subset $c_x, ..., c_y$ of $c_1, ..., c_n$, $\sim (\sim E >\sim (C_x \& ... C_y))$ [ineliminability requirement];
>
> **(iv)** the chain of events $\{x, ...y\}$ between c_1 and e such that $\sim E >\sim Y, ... \sim X >\sim C_1$ (where "between" is to be understood as including c_1 and e in the limit case) must be *unbroken*, and so for each of $c_1, ..., c_n$ [unbroken chain requirement];
>
> **(v)** a chain of events $\{x, ...y\}$ between c and e such that $\sim E >\sim Y, ... \sim X >\sim C$ is broken iff: c is redundant with respect to e such that $\sim (\sim C >\sim E)$; and supposing away redundant events (other than e) which are counterfactually independent of c yet counterfactually related to e — the chain of events between c and e such that $\sim E >\sim Y, ... \sim X >\sim C$ under this supposition would differ from $\{x, ...y\}$ [broken chain definition].

[7] Sartorio approach in one of her papers leaves such a possibility open: "the view that I defend here is not an analysis of causation. It sets a constraint on the concept of cause, and thus it helps to carve up the concept, while at the same time leaving some room for different ways of pinning it down" (Sartorio 2005, 71).

I have argued in the context of various examples that causes satisfy this condition. I have also argued in principle that the Reverse Counterfactual captures our intuitive notion of making a difference. The conclusion of these arguments is that the Reverse Counterfactual can provide a necessary condition for causation, and the above composite condition (i)-(v) is an effort to formalise and concisely state that necessary condition. I have also argued for the following claims concerning it:

- the Reverse Counterfactual captures the difference between causes and mere conditions, by being true of causes but false of mere conditions (where a condition c for e is an event such that $\sim C >\sim E$);
- the Reverse Counterfactual thus offers an account of causal selection, and of the way that it is sensitive to context, by assimilating the principles governing causal selection to the principles governing our context-sensitive assessment of counterfactuals;
- the Reverse Counterfactual thus explains why we use causation to select, so addictively and in such a wide range of contexts (eg. common speech, explanation, prediction, manipulation, ethics and law);
- the Reverse Counterfactual captures the circumstances in which causation fails to be transitive, by being true of causes only when they are sufficiently proximate;
- the Reverse Counterfactual partly explains why causation sometimes appears to be transitive, by making it effectively transitive in certain circumstances (the chain c–d–e will always be transitive when c causes d which causes e, and when d is a condition for e);
- the Reverse Counterfactual distinguishes causes from preempted events, because, although the Reverse Counterfactual may be true of preempted events, the chain from a preempted event to an effect will include events which actually fail to occur;
- the Reverse Counterfactual is true of symmetrically overdetermining causes, allowing us to say that overdetermining causes each cause their effect.

We have, however, seen that a sufficient condition for causation cannot be stated using counterfactuals, due to the impossibility of capturing the metaphysical asymmetry of causation with any temporal asymmetry. When c causes

e, the Reverse Counterfactual is true; but there is more to c causing e than counterfactual dependence between c and e. The Reverse Counterfactual thus offers as much of a counterfactual analysis of causation as can be given.

Bibliography

G.E.M. Anscombe. Causality and extensionality. *The Journal of Philosophy*, 66(6):152–159, 1969.

G.E.M. Anscombe. *Causality and Determination*. Cambridge University Press, Cambridge, 1971.

Helen Beebee. Causing and nothingness. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 291–308. MIT Press, Cambridge, Massachusetts, 2004.

Jonathan Bennett. Counterfactuals and temporal direction. *The Philosophical Review*, 93(1):57–91, 1984.

Jonathan Bennett. On forward and backward counterfactual conditionals. In G. Preyer and F. Siebelt, editors, *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis*, pages 177–202. Rowman and Littlefield, Maryland, 2001.

Jonathan Bennett. *A Philosophical Guide to Conditionals*. Oxford University Press, Oxford, 2003.

Gunnar Björnsson. How effects depend on their causes, why causal transitivity fails, and why we care about causation. *Philosophical Studies*, 133:349–390, 2007.

Thomas D. Bontly. What is an empirical analysis of causation? *Synthese*, 151: 177–200, 2006.

Richard Bradley. A defence of the ramsey test. *Mind*, 116(461):1–21, 2007.

Alex Broadbent. Reversing the counterfactual analysis of causation. *International Journal of Philosophical Studies*, 15:169–189, 2007.

John Collins. Preemptive prevention. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 107–117. MIT Press, Cambridge, Massachusetts, 2004.

John Collins, Ned Hall, and L.A. Paul. *Causation and Counterfactuals*. MIT Press, Cambridge, Massachusetts, 2004.

Phil Dowe. Backwards causation and the direction of causal processes. *Mind*, 105:227–248, 1996.

Phil Dowe and Paul Noordhof. *Cause and Chance: Causation in an Indeterministic World*. Routledge, London, 2004.

Dorothy Edgington. On conditionals. *Mind*, 104:235–329, 1995.

Dorothy Edgington. Counterfactuals and the benefit of hindsight. In P. Dowe and P. Noordhof, editors, *Cause and Chance: Causation in an Indeterministic World*, pages 12–27. Routledge, London, 2004.

Adam Elga. Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science (Proceedings)*, 68:S313–S324, 2000.

Kit Fine. Review: Critical notice. *Mind*, 84(335):451–458, 1975.

Allan Gibbard and William Harper. Counterfactuals and two kinds of expected utility. In *Foundations and Applications of Decision Theory*. Dordrecht, Reidel, 1978.

Nelson Goodman. *Fact, Fiction and Forecast*. Harvard University Press, Cambridge, Massachusetts, fourth edition, 1983.

C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

Paul Grice. Logic and conversation. In *Syntax and Semantics, Volume 3 — Speech Acts*, pages 41–58. Academic Press, London, 1975.

Paul Grice. Further notes on logic and conversation. In *Syntax and Semantics, Volume 9 — Pragmatics*, pages 41–58. Academic Press, London, 1978.

Ned Hall. Causation and the price of transitivity. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 181–204. MIT Press, Cambridge, Massachusetts, 2004a.

Ned Hall. Two concepts of causation. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 225–276. MIT Press, Cambridge, Massachusetts, 2004b.

Ned Hall and L.A. Paul. Causation and preemption. In Peter Clark and Katherine Hawley, editors, *Philosophy of Science Today*, pages 100–130. Oxford University Press, Oxford, 2003.

H.L.A. Hart and A. Honore. *Causation in the Law*. Clarendon Press, Oxford, second edition, 1985.

Daniel Hausman. *Causal Asymmetries*. Cambridge University Press, Cambridge, 1998.

Christopher Hitchcock. The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98:273–299, 2001.

David Hume. *An Enquiry Concerning Human Understanding*. Clarendon Press, Oxford, 1748. This edition: 1902, ed. L. A. Selby-Bigge.

Frank Jackson. *Conditionals*. Blackwell, Oxford, 1987.

Nancy Krieger. Epidemiology and the web of causation: Has anybody seen the spider? *Social Science and Medicine*, 39:887–903, 1994.

Igal Kvart. *A Theory of Counterfactuals*. Hackett Publishing, Indianapolis, 1986.

Igal Kvart. Counterfactuals and causal relevance. *Pacific Philosophical Quarterly*, 72:314–337, 1991.

Igal Kvart. Causal independence. *Philosophy of Science*, 61:96–114, 1994.

David Lewis. Causation. *Journal of Philosophy*, 70:556–567, 1973a. Page numbers refer to (Lewis 1986).

David Lewis. *Counterfactuals*. Harvard University Press, Cambridge, Massachussetts, 1973b.

David Lewis. Counterfactuals and comparative possibility. *Journal of Philosophical Logic*, 2:418–446, 1973c. Page numbers refer to (Lewis 1986).

David Lewis. Counterfactual dependence and time's arrow. *Noûs*, 13:455–476, 1979. Page numbers refer to (Lewis 1986).

David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59: 5–30, 1981.

David Lewis. Putnam's paradox. *Australasian Journal of Philosophy*, 62(3): 221–236, 1984.

David Lewis. *Philosophical Papers, Volume II*. Oxford University Press, Oxford, 1986a.

David Lewis. Causal explanation. In *Philosophical Papers, Volume II*, pages 214–241. Oxford University Press, Oxford, 1986b.

David Lewis. Events. In *Philosophical Papers, Volume II*, pages 214–269. Oxford University Press, Oxford, 1986c.

David Lewis. Humean supervenience debugged. *Mind*, 103:473–490, 1994.

David Lewis. Causation as influence. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 75–106. MIT Press, Cambridge, Massachusetts, 2004a.

David Lewis. Void and object. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 277–290. MIT Press, Cambridge, Massachusetts, 2004b.

Peter Lipton. Causation outside the law. In H. Gross and T.R. Harrison, editors, *Jurisprudence: Cambridge Essays*, pages 127–148. Oxford University Press, Oxford, 1992.

Peter Lipton. Making a difference. *Philosophica*, 51(1):39–54, 1993.

Peter Lipton. Tracking track records. *Proceedings of the Aristotelian Society — Supplementary Volume*, 74(1):179–205, 2000.

Peter Lipton. *Inference to the Best Explanation*. Routledge, London and New York, second edition, 2004.

Barry Loewer. Determinism and chance. *Studies in History and Philosophy of Modern Physics*, 32:609–620, 2001.

John Mackie. Causes and conditions. *American Philosophical Quarterly*, 2: 245–264, 1965.

John Mackie. *The Cement of the Universe*. Oxford University Press, Oxford, 1974.

Michael McDermott. Redundant causation. *The British Journal for the Philosophy of Science*, 46(4):523–544, 1995.

Alice McEleney and Ruth M.J. Byrne. Spontaneous counterfactual thoughts and causal explanations. *Thinking and Reasoning*, 12:235–255, 2006.

Peter Menzies. Is causation a genuine relation? In *Real Metaphysics: Essays in Honour of D.H. Mellor*, pages 120–136. Routledge, London, 2002.

Peter Menzies. Difference-making in context. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 139–180. MIT Press, Cambridge, Massachusetts, 2004.

Peter Menzies and Huw Price. Causation as a secondary quality. *British Journal for the Philosophy of Science*, 44(2):187–203, 1993.

John Stuart Mill. *A System of Logic, Ratiocinative and Inductive*. Harper and Brothers, New York, eighth edition, 1887.

Robert Nozick. *Philosophical Explanations*. Harvard University Press, Cambridge, Massachusetts, 1981.

M. Parascandola and D.L. Weed. Causation in epidemiology. *Journal of Epidemiology and Community Health*, 55:905–912, 2001.

Kit Patrick. Causation by absence. Dissertation for University of Cambridge Natural Sciences Tripos Part II., 2005.

L.A. Paul. Aspect causation. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 205–224. MIT Press, Cambridge, Massachusetts, 2004.

Karl Popper. The arrow of time. *Nature*, 177:538, 1956.

Huw Price. Agency and causal symmetry. *Mind*, 101(403):501–520, 1992a.

Huw Price. The direction of causation: Ramsey's ultimate contingency. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1992:253–267, 1992b.

Huw Price. *Time's Arrow and Archimedes' Point*. Oxford University Press, Oxford, 1996.

Alexander R. Pruss. David lewis's counterfactual arrow of time. *Noûs*, 37: 606–637, 2003.

W.V.O. Quine. *Word and Object*. MIT Press, Cambridge, Massachussetts, 1960.

Murali Ramachandran. A counterfactual analysis of indeterministic causation. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 387–402. MIT Press, Cambridge, Massachusetts, 2004.

Frank Ramsey. General propositions and causality. In Hugh Mellor, editor, *Foundations*, pages 133–151. Humanities Press, Atlantic Highlands, N.J., 1978.

Bertrand Russell. On the notion of a cause. In *Mysticism and Logic*. Allen and Unwin, London, 1917.

Wesley C. Salmon. Probabilistic causality. In *Causation*, pages 137–153. Oxford University Press, New York, 1993.

Carolina Sartorio. Causes as difference-makers. *Philosophical Studies*, 123: 71–96, 2005.

Jonathan Schaffer. Trumping preemption. In J. Collins, N. Hall, and L.A. Paul, editors, *Causation and Counterfactuals*, pages 59–73. MIT Press, Cambridge, Massachusetts, 2004a.

Jonathan Schaffer. Counterfactuals, causal indepedence and conceptual circularity. *Analysis*, 64:299–309, 2004b.

Jonathan Schaffer. Contrastive causation. *Philosophical Review*, 114(3):297–328, 2005.

Jonathan Schaffer. Deterministic chance? *British Journal for the Philosophy of Science*, 58:113–140, 2007.

John Searle. *Speech Acts*. Cambridge University Press, Cambridge, 1969.

John Searle. Indirect speech acts. In *Syntax and Semantics, Volume 3 — Speech Acts*, pages 59–82. Academic Press, London, 1975.

Robert Stalnaker. A defense of conditional excluded middle. In *Ifs*. D. Reidel Publishing Company, Dordrecht, Holland, 1981.

Richard Taylor. *Action and Purpose*. Humanities Press, New York, 1974.

Georg Henrik von Wright. *Causality and Determinism*. Columbia Univerisity Press, New York, 1974.

Caroline Whitbeck. Causation in medicine: The disease entity model. *Philosophy of Science*, 44:619–637, 1977.

Timothy Williamson. *Knowledge and Its Limits*. Oxford University Press, Oxford, 2000.

Timothy Williamson. Knowledge of counterfactuals. 2007.

Stephen Yablo. De facto dependence. *Journal of Philosophy*, 99:130–148, 2002.