

CAUSATION, COUNTERFACTUAL DEPENDENCE AND PLURALISM

by

Francis Longworth

B.A., M.A. Oxford University 1992

Submitted to the Graduate Faculty of
Arts and Sciences in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy.

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

SCHOOL OF ARTS AND SCIENCES

Thus dissertation was presented

by

Francis Longworth

It was defended on

August 17th, 2006

and approved by

Cian Dorr, Assistant Professor, Department of Philosophy

John Earman, University Professor, Department of History and Philosophy of Science

Jim Woodward, Professor, Division of Humanities and Social Sciences, California Institute of
Technology

John Norton, Professor, Department of History and Philosophy of Science

CAUSATION, COUNTERFACTUAL DEPENDENCE AND PLURALISM

Francis Longworth, Ph.D.

University of Pittsburgh, 2006

The principal concern of this dissertation is whether or not a conceptual analysis of our ordinary concept of causation can be provided. In chapters two and three I show that two of the most promising univocal accounts (the counterfactual theories of Hitchcock and Yablo) are subject to numerous counterexamples. In chapter four, I show that Hall's pluralistic theory of causation, according to which there are *two* concepts of causation, also faces a number of counterexamples. In chapter five, I sketch an alternative, broadly pluralistic theory of token causation, according to which causation is a 'cluster concept' with a 'prototypical' structure. This theory is able to evade the counterexamples that beset other theories and, in addition, offers an explanation of interesting features of the concept such as the existence of borderline cases, and the fact that some instances of causation seem to be 'better' examples of the concept than others.

TABLE OF CONTENTS

| | | |
|----------------------|---|-----------|
| PREFACE | | XI |
| 1.0 | AIMS, METHODOLOGY AND OVERVIEW | 1 |
| 1.1. | AIMS | 1 |
| 1.2. | METHODOLOGY | 2 |
| 1.3. | OVERVIEW | 4 |
| 1.3.1. | Chapter Two | 4 |
| 1.3.2. | Chapter Three | 4 |
| 1.3.3. | Chapter Four | 6 |
| 1.3.4. | Chapter Five | 7 |
| 2.0 | THE ACTIVE ROUTE THEORY | 8 |
| 2.1. | INTRODUCTION AND OVERVIEW | 8 |
| 2.2. | NAÏVE DEPENDENCE | 8 |
| 2.3. | THE TRANSITIVITY THESIS | 10 |
| 2.4. | COUNTEREXAMPLES TO TRANSITIVITY | 12 |
| 2.5. | HOLDING FIXED (HF) | 13 |
| 2.6. | ACTIVE CAUSAL ROUTES (H1) | 14 |
| 2.6.1. | Causal Models: Variables, Structural Equations and Causal Graphs | 14 |
| 2.6.2. | Active Causal Route: Definition | 15 |
| 2.7. | (H1) AND COUNTEREXAMPLES TO TRANSITIVITY | 16 |
| 2.7.1. | Late Preemption | 16 |
| 2.7.2. | <u>Dog Bite</u> | 17 |
| 2.7.3. | <u>Boulder</u> | 19 |
| 2.8. | COUNTEREXAMPLES TO THE SUFFICIENCY OF (H1): | |

| | | |
|---------|--|----|
| | SWITCHING AND SELF-CANCELING THREATS..... | 22 |
| 2.8.1. | First Counterexample to the Sufficiency of (H1): <u>Two Trolleys</u> | 22 |
| 2.8.2. | Second Counterexample to the Sufficiency of (H1): <u>Two Assassins</u> | 26 |
| 2.9. | COUNTERFACTUAL DEPENDENCE AND THE PRINCIPLE OF SUFFICIENT REASON (H3) | 29 |
| 2.10. | (H3), PREEMPTION AND SELF-CANCELING THREATS..... | 32 |
| 2.11. | COUNTEREXAMPLES TO (H3) | 35 |
| 2.11.1. | First Counterexample to (H3) in a Self-Contained Model: <u>Main Generators</u> | 36 |
| 2.11.2. | Second Counterexample to (H3) in a Self-Contained Model: <u>Birth and Death</u> | 38 |
| 2.11.3. | First Counterexample to (H3) in a Non-Self-Contained Model: <u>HAL</u> | 39 |
| 2.11.4. | Second Counterexample to (H3) in a Non-Self-Contained Model: <u>Two Trolleys</u> | 41 |
| 2.12. | (H1) AND (H3) AS MUTUALLY-REINFORCING INFLUENCES ON CAUSAL JUDGMENTS..... | 42 |
| 2.13. | CONCLUSION..... | 43 |
| 3.0 | THE DE FACTO DEPENDENCE THEORY..... | 44 |
| 3.1 | INTRODUCTION AND OVERVIEW..... | 44 |
| 3.2 | DE FACTO DEPENDENCE..... | 45 |
| 3.3 | DE FACTO DEPENDENCE AND CAUSAL MODELS..... | 47 |
| 3.4 | EARLY AND LATE PREEMPTION..... | 49 |
| 3.5 | SELF-CANCELING THREATS..... | 52 |
| 3.6 | SWITCHING | 53 |
| 3.7 | OBJECTIONS TO THE SUFFICIENCY OF (DF) | 54 |
| 3.7.1 | First Counterexample to the Sufficiency of (DF): <u>Flip 1cm</u> | 55 |
| 3.7.2 | Second Counterexample to the Sufficiency of (DF): <u>Two Assassins</u> | 60 |
| 3.7.3 | Third Counterexample to the Sufficiency of (DF): <u>Refusal to</u> | |

| | | | |
|------|-------|--|-----|
| | | <u>Shoot</u> | 62 |
| 3.8 | | OBJECTIONS TO THE NECESSITY OF (DF)..... | 63 |
| | 3.8.1 | First Counterexample to the Necessity of (DF): <u>Double Backup</u> | 63 |
| | 3.8.2 | Second Counterexample to the Necessity of (DF): <u>Nudge</u> | 69 |
| | 3.8.3 | Third Counterexample to the Necessity of (DF): <u>Small Meteorite and Suzy</u> | 71 |
| | 3.8.4 | Fourth Counterexample to the Necessity of (DF): <u>Billy and Suzy Deflect</u> | 75 |
| 3.9 | | (DF) IS UNDECIDABLE..... | 77 |
| | 3.9.1 | The Counterpart Problem..... | 77 |
| | 3.9.2 | The Halting Problem..... | 79 |
| 3.10 | | CONCLUSIONS..... | 80 |
| 4.0 | | PLURALISTIC THEORIES OF CAUSATION..... | 81 |
| | 4.1 | INTRODUCTION AND OVERVIEW..... | 81 |
| | 4.2 | UNCONTROVERSIAL TYPES OF PLURALISM..... | 82 |
| | 4.2.1 | Plurality of Causes of an Event..... | 82 |
| | 4.2.2 | Type/Token Plurality..... | 82 |
| | 4.3 | HITCHCOCK'S NET AND COMPONENT CAUSES..... | 83 |
| | 4.3.1 | The Deterministic Case..... | 86 |
| | 4.3.2 | The Indeterministic Case..... | 88 |
| | 4.4 | HALL'S TWO CONCEPTS OF CAUSATION (TC)..... | 92 |
| | 4.4.1 | Preemption, Prevention and Omission..... | 93 |
| | 4.4.2 | Counterexamples to (TC) | 95 |
| | 4.5 | CONCLUSION..... | 98 |
| 5.0 | | THE PROTOTYPE THEORY OF CAUSATION..... | 99 |
| | 5.1 | INTRODUCTION AND OVERVIEW..... | 99 |
| | 5.2 | FIVE THESES AND THREE DESIDERATA..... | 100 |
| | 5.3 | RESEMBLANCE TO PARADIGM THEORIES..... | 107 |
| | 5.4 | LAKOFF'S TWELVE PROTOTYPICAL PROPERTIES..... | 109 |
| | 5.5 | CAUSATION AS A CLUSTER CONCEPT..... | 111 |
| | 5.6 | IN FAVOR OF THE CLUSTER CONCEPT..... | 117 |

| | | |
|-----|-------------------|-----|
| 5.7 | CONCLUSION..... | 122 |
| 6.0 | CONCLUSIONS..... | 123 |
| | BIBLIOGRAPHY..... | 125 |

LIST OF TABLES

| | |
|---|----|
| Table 1. Theoretical Verdicts of (H1) and (H3)..... | 42 |
| Table 2. Theoretical Verdicts of (DF)..... | 59 |

LIST OF FIGURES

| | |
|--|----|
| Figure 2.1. <u>Trainee and Supervisor</u> | 15 |
| Figure 2.2. <u>Billy and Suzy</u> | 17 |
| Figure 2.3. <u>Dog Bite (3-variable)</u> | 18 |
| Figure 2.4. <u>Dog Bite (4-variable)</u> | 18 |
| Figure 2.5. <u>Boulder (3-variable)</u> | 19 |
| Figure 2.6. <u>Boulder (4-variable)</u> | 20 |
| Figure 2.7. <u>Model Appropriateness Criteria</u> | 22 |
| Figure 2.8. <u>Flip</u> | 23 |
| Figure 2.9. <u>Flip (3-variable)</u> | 24 |
| Figure 2.10. <u>Two Trolleys</u> | 26 |
| Figure 2.11. <u>Two Assassins</u> | 27 |
| Figure 2.12. <u>Trainee and Supervisor</u> | 33 |
| Figure 2.13. <u>Billy and Suzy</u> | 34 |
| Figure 2.14. <u>Two Assassins</u> | 34 |
| Figure 2.15. <u>Main Generators</u> | 36 |
| Figure 2.16. <u>Birth and Death</u> | 38 |
| Figure 2.17. <u>HAL</u> | 40 |
| Figure 2.18. <u>Two Trolleys</u> | 41 |
| Figure 3.1. <u>Trainee and Supervisor: Annotated Causal Graph</u> | 50 |
| Figure 3.2. <u>Billy and Suzy: Annotated Causal Graph</u> | 51 |
| Figure 3.3. <u>Two Assassins: Annotated Causal Graph</u> | 52 |
| Figure 3.4. <u>Flip: Annotated Causal Graph</u> | 54 |
| Figure 3.5. <u>Trainee and Meteorite: Annotated Causal Graph</u> | 58 |
| Figure 3.6. <u>Two Assassins Revisited: Annotated Causal Graph</u> | 61 |

Figure 3.7. Refusal to Shoot: Annotated Causal Graph.....62

Figure 3.8. Double Backup: Annotated Causal Graph.....65

Figure 3.9a. Triple Backup: Annotated Causal Graph.....67

Figure 3.9b. Triple Backup: Second and Third-Level Fallback Scenarios.....68

Figure 3.10. Nudge: Annotated Causal Graph.....70

Figure 3.11. Small Meteorite and Suzy: Annotated Causal Graph.....72

Figure 3.12. Swinging Bottle: Annotated Causal Graph.....74

Figure 3.13. Billy and Suzy Deflect: Annotated Causal Graph.....76

Figure 3.14. Power Failure: Annotated Causal Graph.....78

Figure 4.1. Birth Control Pills.....86

Figure 4.2. Birth Control Pills (Deterministic Case)87

PREFACE

I wish to thank Peter Bokulich, Cian Dorr, Phil Dowe, John Earman, Matthias Frisch, Peter Godfrey-Smith, Ned Hall, Caspar Hare, Chris Hitchcock, Simon Keller, John Norton, Wendy Parker, Andrea Scarantino, Eric Swanson, Judy Thomson, Jim Woodward and Steve Yablo for helpful discussions of the material in this dissertation.

1.0 AIMS, METHODOLOGY AND OVERVIEW

1.1 AIMS

The overarching question that this dissertation addresses is whether or not a conceptual analysis of our ordinary concept of causation be given. Is it possible to provide a philosophical theory that will deliver the intuitively correct verdicts about when the concept applies and when it does not? There are many different kinds of analysis that one might attempt to provide: one might offer a *descriptive* account, whose aim is to accord extremely closely with our intuitions; or, one might attempt, for some particular purpose, to provide a *revisionary* account. In the latter case, it is less important that the theory follow *all* of the contours of our intuitive causal judgments. I shall be primarily concerned here with descriptive analysis. I do not think that this is the *only* sort of analysis that should be pursued, nor even necessarily the most important or worthwhile. Revisionary accounts that have significant payoffs in terms of, for example, aiding causal inferences, playing a central role in some larger metaphysical system, or more cleanly analyzing other concepts, are of great value, and may be well worth the price of some small divergence from ordinary usage. My perspective is this: if we *do* decide to undertake the project of providing a descriptive conceptual analysis of our ordinary concept of causation, what sorts of difficulties should we expect to come across, and why? How feasible is the project? Are the prospects for developing a *univocal* account good, or should we instead pursue a more pluralistic strategy? If so, what kind of pluralistic theory should we seek?

I will be concerned with the analysis of singular or *token* causation, rather than with general or *type*-level causation. Furthermore, I shall focus principally on the deterministic case, although I will at intervals have something to say about irreducible indeterministic token causation. Finally, I will in general take the relata of the causal relation to be *events*, as is fairly common practice.

In pursuing a project that gives significant weight to intuitions, one is sometimes asked whether one is doing *psychology*, where the primary concern is giving an account of the factors the individuals take into account in forming their intuitive judgments, or *metaphysics*, where the goal is to describe some objective feature of the world that is independent of the vagaries of human judgment. This is a difficult and subtle issue. In the first instance, my focus is on individuals' *intuitions*. Now it may be that those intuitions *do* pick out some reasonably objective feature of the world. But they may not do so. It may be that our intuitive judgments about whether or not one event is a cause of another are influenced by moral and intentional considerations, which may be of a partly subjective nature. If this is so, one may decide that one would prefer a revisionary metaphysics that is more objective and more consistent with other metaphysical commitments one may not be willing to give up. If subtle differences in our intuitions do not correspond to any deep differences at the level of basic ontology, should we care if our theoretical verdicts do not always accord with our intuitions?

A further issue concerns whether a theory of causation should aim to give a *psychologically plausible* account of our actual intuition-forming processes; the steps that we actually carry out in our heads (perhaps only implicitly) in terms of the factors we pay attention to, and how we weigh them up, in coming to our intuitive verdicts. If one takes this to be an important aim, then overly-sophisticated technical theories will seem less plausible candidates.

1.2 METHODOLOGY

The dominant methodological strategy I shall follow in this dissertation will be comparing the verdicts of a variety of theories of causation with those of intuition. When these are not in agreement, it counts against the theory. In chapters two to four, I test some of the leading extant theories of causation against a variety of candidate counterexamples. There is already a canonical set of such cases that is steadily accumulating as the causation literature grows, and I will make frequent use of these canonical cases. However, I also construct a large number of *new* examples, explicitly for the purpose of testing some aspect or other of a particular theory.

In a somewhat unconventional fashion, I give considerable weight to *indeterminate* intuitions; cases in which we find it difficult to say whether some putative C really does cause E.

Rather than sweeping these intuitions under the carpet as unsuitable test cases, I treat them as interesting data in their own right, which a theory should be able to give some account of. David Lewis, in his influential paper ‘Causation’ (1973) states, with regard to cases of symmetric overdetermination, “For me these are useless as test cases because I lack firm naïve opinions about them” (1973). In his later “Postscripts to ‘Causation’” (1986) he adds

If an analysis of causation does not deliver the common-sense answer, that is bad trouble. But when common sense falls into indecision or controversy, or when it is reasonable to suspect that far-fetched cases are being judged by false analogy to commonplace ones, then theory may safely say what it likes. Such cases can be left as spoils to the victor, in D.M. Armstrong’s phrase. We can reasonably accept as true whatever answer comes from the analysis that does best on the clearer cases (p.194).

But Lewis goes on to say, however, (and this is less well-known) that

It would be better still, however, if theory itself went indecisive about the hard cases. If an analysis says that the answer for some hard case depends on underdescribed details, or on the resolution of some sort of vagueness, that would explain nicely why common sense comes out indecisive. (1986).¹

Following Rawls, it has been common for theorists to seek some sort of ‘reflective equilibrium between their theories and their intuitions. I shall not have too much to say about that, but I will insist that individuals must achieve some equilibrium in their intuitive verdicts about cases that they themselves admit are analogous. If they judge in the one case that some putative is a cause of E, then they must do so in the other. Not all intuitions are to be blindly accepted at face value. Some intuitions are *clearly* mistaken, and are corrigible. No theory of causation should be held hostage to mistaken intuitions (unless the goal is specifically to develop a theory of *error*; an account of why it is that people sometimes go astray or are misled in making their causal judgments).

At this point I must introduce a major methodological *caveat*. The data against which I test the various philosophical theories are primarily my *own* intuitions about particular test cases. Secondly, they are the intuitions gleaned from informal surveys of my colleagues, friends and family, and lastly, those intuitions about canonical examples that are generally taken to be correct in the philosophical literature on causation. There is very little in the way of systematic

¹Hiddleston (2005, p.51-52) makes a similar point.

empirical data on what the intuitions of ordinary people actually are. This is a puzzling state of affairs from a methodological point of view; one that one hopes will soon be rectified.

1.3 OVERVIEW OF CONTENT

1.3.1 Chapter Two

In chapter two I begin by discussing a variety of attempted solutions to what has been the major thorn in the side of counterfactual analyses of causation; the fact that effects do *not* depend on their causes in cases of redundant causation (the so-called ‘Preemption Problem’). While these approaches have a good deal of intuitive appeal, I show that all of them are unsuccessful. I begin by showing how Lewis’s appeal to the supposed transitivity of causation run into trouble. I spend the rest of the chapter and chapter three critiquing a what is perhaps *the* most significant theoretical move in the counterfactual analysis of causation: the strategy of looking for counterfactual dependence between putative cause and effect while *holding fixed* certain facts. This approach rejects transitivity and hence avoids all of its attendant problems. However, this Holding Fixed strategy generates two horrible new classes of counterexample: what I call ‘Switching’ and ‘Self-Canceling Threats.’ I first present Hitchcock’s technical formulation of this holding fixed strategy within a causal modeling framework of structural equations and causal graphs: the ‘Active Route Theory (H1).’ Hitchcock introduces of the notion of an *appropriate* causal model, and appeals to the *Principle of Sufficient Reason* in a further iteration (H3) in an attempt to get around the Switching and Self-Canceling Threats counterexamples. But these maneuvers are shown to be unsuccessful and to give the wrong results even for simple cases of preemption.

1.3.2 Chapter Three

In the next chapter, I discuss another impressive technical attempt to solve the Preemption Problem within a counterfactual framework, using the Holding Fixed strategy. Yablo’s *De Facto*

Dependence (DF) theory of causation (2002, 2004) uses the notion of *artificiality* to handle the dreaded cases of Switching and Self-Canceling Threats. The principal idea of Yablo's account is that some of the hidden dependencies that are revealed by holding fixed some fact G have an "artificial" quality, in virtue of which they should *not* be taken to indicate causation. The hope is that the notion of artificiality will be able to distinguish between the latent dependencies in preemptions and those in Switching and Canceled threats, treating the former as genuinely causal, and the latter as artificial.

I present several counterexamples to (DF), attempt a few obvious repairs but conclude that (with one pleasing exception) they do not work. (DF) is able to handle multiple redundancy but its technical machinery is so complex and so sprawling that, while it delivers the correct results for some very tricky cases, it then comes unstuck with the much simpler cases.

I also show that (DF) has the potentially worrying feature of not always delivering determinate verdicts on whether or not event C causes event E. But this is not the good kind of indeterminacy generating that lines up nicely with the indeterminacy in our intuition. The theory appears undecidable for cases that are intuitively very clear.

I view the introduction of the notions of *artificiality*, of an *appropriate* causal model and the appeal to the *Principle of Sufficient Reason* as the addition of further epicycles to an already degenerating research program. If the goal of a philosophical theory of causation is taken to be the descriptive analysis of our ordinary concept, then I suggest that counterfactual theories should be abandoned.

Other univocal accounts, do not fare any better than counterfactual theories. Lewis (1973) presents objections to earlier regularity theories of causation, Eells (1991) presents objections to probabilistic theories and Woodward (and especially Schaffer (2005)) present objections to physical process theories of Salmon and Dowe. These objections are well-known. Given the difficulties with (H1)–(H3) and (DF), perhaps the most promising of all the univocal analyses of causation, one is inclined to think that the prospects for finding a successful univocal theory, whether counterfactual or otherwise rather bleak. Obviously, the failure of n univocal theories does not entail that theory $n+1$ will fail. But it is indicative, however, of the magnitude of the task facing the univocalist.

One way forward would be to abandon the search for a univocal theory. In chapter four, I discuss two pluralistic treatments of causation that have been offered, which are contrasting types.

1.3.3 Chapter Four

Recently, Hall (2004), Hitchcock (2003) and Cartwright (1999, 2004, forthcoming) have offered pluralistic accounts of causation. The idea that there is a plurality of types of causation, however, is not new. It goes back at least as far as Aristotle's distinction between material, formal, efficient and final causes, and can also be found in Spinoza and Hume. There are a variety of ways in which one might be a pluralist about causation. In this chapter, however, I focus on those varieties of pluralism that *preclude the possibility of a univocal conceptual analysis* of causation, and mention the others only briefly.

I begin with a brief discussion of some *uncontroversial* forms of pluralism about causation that are not terribly interesting, which do *not* threaten the univocalist. In section 4.3, I discuss a form of pluralism that *is* a threat to the univocalist. Hitchcock (2003) claims that individuals disagree with one another about whether or not C is a cause of E, and that this disagreement is much more widespread than has previously been supposed. If two individuals disagree, he argues, and neither is *obviously* mistaken, then no univocal theory could possibly accommodate both intuitions, since the intuitions are incompatible. I discuss one particular case, Birth Control Pills, which Hitchcock uses to argue that a distinction should be made between 'net' and 'component' causes. No single analysis, it is claimed, can (or should attempt to) accommodate both subconcepts. Instead, each requires a separate analysis. I will argue that while it is useful for certain practical purposes to introduce, and provide stipulative definitions for the notions of net and component cause, *we don't need to do so* for the purposes of analyzing our ordinary concept of token causation.

I next discuss Hall's extremely interesting and creative pluralistic theory of causation (Hall 2004). Hall proposes that there are *two* concepts of causation: 'production' and 'dependence', each requiring a different analysis. Dependence is just counterfactual dependence; production is a local and intrinsic relation, perhaps involving nomic sufficiency, although a full

analysis is not given. Hall's definition is disjunctive: C is a cause of E if and only if C produces E or E depends on C. Both disjuncts are sufficient for C to be a cause of E. This dualistic theory is able to deal with early and late preemption in a psychologically natural (and therefore plausible) manner. Preempting causes count as *bona fide* causes not in virtue of any *dependence* of the effect on the cause (while holding fixed the redundant backup, as the counterfactualist would have it), but rather in virtue of the local productive relation between the cause and the effect. Hall's pluralistic theory, since it does not need to appeal to the Holding Fixed maneuver, has the great advantage of not ruling in Switching and Self-Canceling Threats as genuine cases of causation. This is a very important advance on univocal theories. Unfortunately, there are quite a few counterexamples to Hall's theory: cases that we intuitively judge to be causation, but which exhibit neither production nor dependence, and cases that exhibit both production and dependence, yet which are arguably not causation. We therefore have reasons to suspect that if causation *is* a non-univocal concept, Hall's non-univocal analysis is not quite the right one.

1.3.4 Chapter Five

In this chapter I offer an alternative pluralistic theory, which has a Wittgensteinian flavor. I suggest that our concept of causation has a *prototype* structure. I suggest that there are a large number or 'cluster' of properties that we take to be relevant to our causal judgments, none of which are individually necessary, but various combinations of which are sufficient. There are a variety of related senses of the concept, which do not share any set of individually necessary and jointly sufficient conditions. Causation then defies univocal analysis in those terms. In addition to handling problematic counterexamples in a superior fashion to other theories, this cluster theory explains several features of the concept that would be difficult to account for, if the concept had a classical necessary and sufficient conditions structure. The theory is also attractive in that it appears to be empirically testable to some degree, in particular, with regard to the existence of so-called 'prototype effects', which were first recognized by Eleanor Rosch in the 1970s, and constitute a major challenge to conventional conceptual analysis.

2.0 THE ACTIVE ROUTE THEORY OF CAUSATION

2.1 INTRODUCTION AND OVERVIEW

In this chapter I discuss a variety of attempted solutions to what has been the greatest difficulty facing counterfactual theories of causation; the fact that effects do *not* depend on their causes in cases of redundant causation (the so-called ‘Preemption Problem’). While these attempted solutions are promising and intuitively well-motivated, I show that all of them are unsuccessful. In section 2.2, I begin by presenting a naïve theory of counterfactual dependence, and show how it falls to simple cases of preemption. In section 2.3, I briefly discuss Lewis’s early attempts to deal with such cases by introducing the thesis that causation is a *transitive* relation. In section 2.4 I present some well-known counterexamples to Lewis’s theory. In section 2.5 I outline an alternative strategy for dealing with the Preemption Problem (that does not assume transitivity), which involves searching for hidden counterfactual dependence by ‘holding fixed’ certain facts. In sections 2.6-2.7, I present Hitchcock’s technical formulation of this strategy within a causal modeling framework: the ‘Active Route Theory’ (H1). I show in section 2.8 that (H1) is subject to a variety of uncontroversial counterexamples involving ‘Switching’ and ‘Self-Canceling Threats.’ In sections 2.9-2.11, I criticize a new variant of this approach, also due to Hitchcock, (H3), and show that in addition to being subject to many of the same counterexamples as (H1), (H3) cannot deliver the intuitively correct verdicts for simple cases of preemption. Finally, in section 2.12, I show that a combined version of (H1) and (H3) also fails.

2.2 NAÏVE DEPENDENCE

David Hume, in the *Enquiry Concerning Human Understanding* (1748), pointed to a link between causation and counterfactual dependence:

[W]e may define a cause to be *an object followed by another, and where all the objects, similar to the first are followed by objects similar to the second*. Or, in other words, *where, if the first object had not been, the second never had existed*. (1748, Section VII, Part II).²

Following Hume, an initial counterfactual analysis (call it Naïve Dependence) might be formulated as:

(ND) C is a cause of E if and only if E *counterfactually depends* on C. In other words, if C had not occurred, E would not have occurred.³

It is well known, however, that effects do not always depend counterfactually on their causes. Consider:

Trainee and Supervisor: Trainee and Supervisor are on a mission to kill Victim. Trainee shoots first and Victim bleeds to death. Supervisor, observing that Trainee has pulled the trigger, does not shoot. If Trainee hadn't shot, however, Supervisor would have stepped in and done so, again resulting in Victim's bleeding to death.⁴

Although Victim's bleeding to death would have depended on Trainee's shooting in the absence of Supervisor, Supervisor's presence breaks this dependence. Such cases, in which the actual cause *preempts* some redundant backup, are known as cases of 'preemption'. Preemption therefore presents a problem for those who wish to base an account of causation on counterfactual dependence. Call the lack of dependence in cases of preemption the "Preemption Problem". The Preemption Problem has caused huge difficulties for the counterfactual analyst of causation; indeed much of the literature on counterfactual theories concerns attempts to get around the Preemption Problem, by adding further conditions that deliver the correct theoretical verdict that preemption *is* causation, hopefully without introducing any new counterexamples.

² Hume's italics.

³ Lets us, in keeping with common practice, take the relata in (ND) to be *events* rather than Hume's 'objects'.

⁴ Adapted from Hitchcock (2001)).

Let us look at the first significant attempt to solve the Preemption Problem: David Lewis's appeal to the thesis that causation is always *transitive*.

2.3 THE TRANSITIVITY THESIS

Lewis's counterfactual theory, presented in his seminal "Causation" (1973), was the first real advance on (ND). The theory relied heavily on the assumption that the causal relation is transitive. Call this thesis 'Transitivity'.

Transitivity: Causation is transitive if and only if:

If C causes D and D causes E, C is a cause of E.

It seems intuitively plausible that causation is transitive: think of a line of dominoes toppling one after the other: the first causes the second to fall, the second causes the third to fall, and it seems correct to say that the first domino is also a cause of the third domino's falling. It doesn't seem unreasonable to expect that transitivity would hold generally. In fact, Transitivity may be one of our central platitudes about causation.

Lewis's analysis of causation, for the purposes of this chapter, can be summarized as:

(L) C is a cause of E if and only if there is a chain of intermediate events $D_1 \dots D_n$ between C and E such that E counterfactually depends upon D_n , D_n counterfactually depends upon D_{n-1} , ... and D_1 counterfactually depends upon C.⁵

Lewis states that counterfactuals must not *backtrack*. If we are considering a world in which some event D_n in a chain of dependency $D_1 \dots D_n$ did not occur, D_{n-1} would still have occurred; so too would D_{n-2} , and all the other intermediate events stretching back to (and including) C. We are to understand the non-occurrence of D_n , Lewis says, as a "minor miracle": D_n is to be cleanly excised from the causal history of E, with no disruption to prior events.

⁵ The truth conditions of counterfactuals are given in terms of a similarity metric of possible worlds to the actual world. The technical details of the account need not concern us here.

Let us see how introducing Transitivity enables (L) to get the right result for our case of preemption, Trainee and Supervisor. Trainee's shot is linked to Victim's dying by a chain of dependence: Victim's dying depends counterfactually on some intermediate event, for example, the flight of the bullet at some particular intermediate point en route to Victim (call this event 'B'). B in turn depends on Trainee's firing. If Trainee had not shot, B would not have occurred; if B had not occurred, Victim would not have died. Note that if B had not occurred, *Trainee would still have shot* (the 'no backtracking' rule). Hence Trainee's firing causes B and B causes Victim's death. Invoking Transitivity, Trainee caused Victim to die. Does this solve the Preemption Problem? Unfortunately not. Lewis's theory falls to cases of "late preemption":

Billy and Suzy: Billy and Suzy each throw a rock at a bottle. Suzy's arrives first and the bottle shatters. Billy's rock arrives a split-second later, encountering only flying shards of glass.

It is intuitively obvious that Suzy's rock rather than Billy's caused the bottle to shatter, but in this case, there is neither simple counterfactual dependence, nor a chain of counterfactual dependence from Suzy's throw to the bottle's shattering. In contrast to early preemption, we *cannot* say that if Suzy's rock had not been at some intermediate position en route to the bottle the bottle would not have shattered, because *Billy would still have thrown*. The fact that Billy throws is independent of Suzy's throw. Hence we cannot use the 'no backtracking' rule to ensure that Billy doesn't throw.

In Trainee and Supervisor, the backup process is cut short by C, early on. In Billy and Suzy, however, the backup process (the approach of Billy's rock) is only terminated very late, by the occurrence of the effect E itself. For this reason, the two cases are referred to as *early* and *late* preemption respectively. Late preemption counterexamples stalled the counterfactual research program for many years. Lewis (2000, 2004) himself responded to late preemption by redefining causation as the ancestral of "influence", a more fine-grained version of counterfactual dependence. I will not go into the details here, but as with his earlier theory, his new theory also relies on Transitivity.

2.4 COUNTEREXAMPLES TO TRANSITIVITY

Transitivity, however, has some undesirable consequences. There seem to be counterexamples in which C causes D, and D causes E, but C is *not* a cause of E. *New* counterexamples arise that the naïve theory (ND) was able to handle correctly. Consider the following two examples:

Dog Bite: Terrorist is about to push the detonator button with his right hand. Dog bites off his right hand, so Terrorist uses his left hand to push the button. An explosion ensues.

Boulder: A large boulder dislodges and falls towards Hiker. Hiker sees the boulder and ducks. Hiker survives.

We do not intuitively feel that the dog's biting Terrorist's hand was a cause of the explosion. Nor do we think that the boulder's falling was a cause of the Hiker's survival. Note that (ND) delivers the intuitively correct theoretical verdict for both of these cases: the bomb's explosion does not depend on the dog's biting, and the Hiker's survival does not depend on the boulder's falling.

What does (L) say about these cases? In the first example, the dog's biting off the Terrorist's right hand caused him to push the button with his left hand, and his pushing the button with his left hand caused explosion. Transitivity therefore requires that the dog's biting caused the explosion. Similarly, the falling of the boulder causes Hiker to survive, according to (L). Both results are highly counterintuitive.

Boulder is an instance of what I shall call a "Self-Canceling Threat". Self-Canceling Threats have the following structure: C introduces some threat to E, but at the same time also initiates some countermove that is successful in canceling the threat to E, and E subsequently occurs. In Boulder, the boulder's falling poses a threat to Hiker's survival, but at the same time, the motion of the boulder alerts Hiker to its approach. Hiker takes evasive action thus canceling the threat to his survival.

One response to Dog Bite and Boulder would be simply to reject Transitivity. But doing so would forfeit the solution to the Preemption Problem. What is required is an alternative

solution to the Preemption Problem that doesn't generate counterexamples such as Dog Bite and Boulder.

Notice that in the two cases of preemption above (Trainee and Supervisor and Billy and Suzy), while the effects do not depend on their causes, they *do* depend on them if we *hold fixed* certain facts. Victim's bleeding to death *does* depend on Trainee's shooting, if we hold fixed the fact that Supervisor doesn't actually fire. Similarly, the bottle's shattering *does* depend on Suzy's throwing, if we hold fixed the fact that Billy's rock does not hit the intact bottle. By holding the right facts fixed, we are able to reveal the latent dependencies between cause and effect that are hidden by the presence of the preempted backups.

This Holding Fixed strategy to the Preemption Problem is attractive because it does not make any reference to the problematic Transitivity thesis, which opened up the counterfactual theory to Dog Bite and Boulder counterexamples. Let us try to formalize this approach.

2.5 HOLDING FIXED (HF)

One simple candidate formulation of a "holding-fixed" counterfactual theory is:

(HF) C is a cause of E if and only if E counterfactually depends on C holding fixed some fact G.

(HF) bears a close relation to familiar epistemic practices for discovering the causes of natural phenomena. The Galilean notion of experiment involves trying to reveal causal relationships by manipulating some candidate cause and looking for the anticipated effect, while screening off any potentially interfering factors by holding them fixed. (HF) is currently a popular strategy in the causation literature; accounts giving a central place to some version of (HF) have been proposed, most notably, by Hitchcock (2001) and Yablo (2002, 2004), and also by Pearl (2000), Halpern and Pearl (2001, 2005) and Woodward (2003).

2.6 ACTIVE CAUSAL ROUTES (H1)

Hitchcock develops an impressively sophisticated version of (HF) within a causal modeling framework of structural equations and causal graphs. He relies heavily on the techniques of Judea Pearl, which were developed primarily for *type* rather than *token* causation, for the purposes of casual inference. Hitchcock provides a counterfactual construal of equations and graphs suitable for their application to the analysis of token causation.

2.6.1 Causal Models: Variables, Structural Equations and Causal Graphs

The following is a very condensed introduction to equations and graphs and Hitchcock's counterfactual construal of them. The reader is referred to Hitchcock (2001, p.279-287) for more comprehensive details. I will work through the Trainee and Supervisor and Billy and Suzy examples and try to make the principles intuitive.

A *causal model* consists of a series of *variables* together with an accompanying set of *structural equations*, which encode the various relations of counterfactual dependence that hold between those variables. Let us illustrate the construction of a causal model for Trainee and Supervisor. When modeling this case, it is natural to take the events "Trainee shoots", "Supervisor shoots" and "Victim bleeds to death" as the variables (call them T, S, and V). These variables are binary, and are ascribed the value 1 if they occur and 0 if they do not. In the actual world, Trainee shoots, so $T=1$. The full set of Boolean structural equations is then: $T=1$, $S=\sim T$, $V=T\vee S$. Note that independent variables go on the right side of the equality sign, dependent variables on the left. In the actual world, the equations yield: $T=1$, $S=0$, $V=1$. Counterfactually, if Trainee *hadn't* shot: $T=0$, $S=1$, $V=1$. If they had *both* shot, $V=1$; if neither had shot, $V=0$.

The structural equations can be given a compact visual representation as a *causal graph* (figure 2.1). Variables are linked by directed arrows. If a variable Y depends on another variable X (as indicated in the structural equations), an arrow is drawn from X to Y. "Negative" relationships may be indicated with a '-' sign adjacent to the relevant arrow. I suggest the following additional notation, which greatly facilitates the evaluation of the theoretical verdict of (H1) in particular cases. The value a variable takes in the actual world is written *above* the

variable, and indicated by an ‘ α ’ (for ‘actual’); the value V would have taken counterfactually, if C had not occurred, is written *below* the variable, and indicated by ‘CF’ (for counterfactual). In each of the figures in this chapter, I have included the theoretical verdict for each case (e.g. (H1)=Yes), together with our intuitive judgment (e.g. Intuition=No).

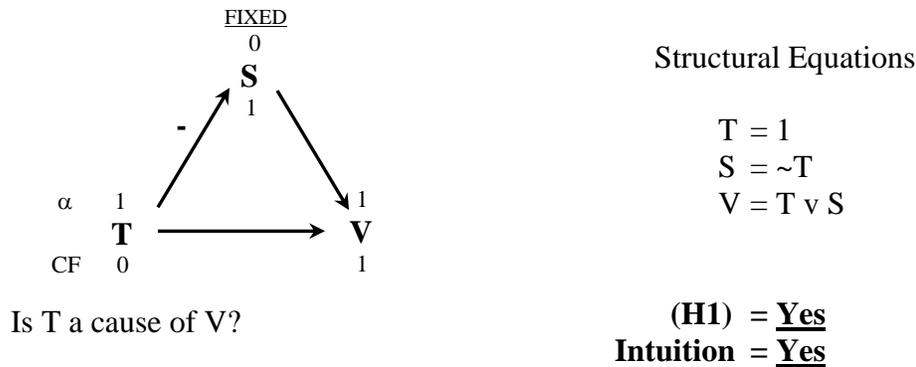


Figure 2.1. Trainee and Supervisor

2.6.2 Active Causal Route: Definition

A *route* is simply an ordered n -tuple of variables, the ordering being determined by the direction of the arrows. For example: $\langle T, S, V \rangle$ or $\langle T, V \rangle$ in figure 2.1. A route $\langle X, Y_1, \dots, Y_n, Z \rangle$ is *active* if and only if Z counterfactually depends on X , with all the variables that are *not* on the route held fixed at their actual values. $Y_1 \dots Y_n$ are *not* held fixed, but are allowed to vary in accordance with their respective structural equations. Hitchcock’s formal definition is:

(ACT): The route $\langle X, Y_1, \dots, Y_n, Z \rangle$ is active in the causal model $\langle V, E \rangle$ if and only if Z depends counterfactually upon X within the new system of equations E' constructed from E as follows: for all $Y \in E$ if Y is intermediate between X and Z but does not belong to the route $\langle X, Y_1, \dots, Y_n, Z \rangle$, then replace the equation for Y with a new equation that sets Y equal to its actual value in E . (If there are no intermediate variables that do not belong to this route then E' is just E). (2001, p. 286)

Causation is then defined:

(H1) Let C and E be distinct occurrent events, and let X and Z be variables such that the values of X and Z represent alterations [in value] of C and E respectively. Then C is a cause of E if and only if there is an *active causal route* from X to Z in an appropriate causal model.

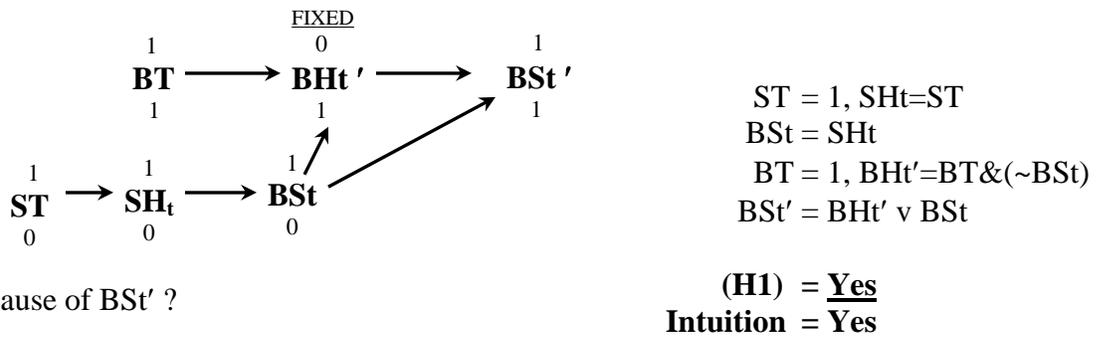
What does (H1) say for Trainee and Supervisor? It is readily seen that $\langle T, V \rangle$ is an active route. Fixing S at its actual value (i.e. $S=0$), V counterfactually depends on T: for when $T=0$, $V=0$; and when $T=1$, $V=1$. Hence, our causal model of Trainee and Supervisor delivers the intuitively correct verdict that Trainee's shooting *is* a cause of Victim's bleeding to death.

2.7 (H1) AND COUNTEREXAMPLES TO TRANSITIVITY

By doing without transitivity, Hitchcock's theory offers the promise of handling early and late preemption correctly, without generating the counterexamples (Dog Bite and Boulder) that the Transitivity thesis incurred for Lewis' theory (L).

2.7.1 Late Preemption

Late preemption can be handled in a similar manner to early preemption. Recall that it is the bottle's shattering that terminates the backup process. Let us take our causal model to include the binary variables Suzy Throws, Billy Throws, Suzy's rock Hits at time t, Billy's Hits at t' (fractionally later than t), the Bottle's being shattered at t and Bottle's being shattered at t' (ST, BT, SHt, BHt', BSt and BSt' respectively). The structural equations and associated causal graph are:



Is ST a cause of BSt' ?

Figure 2.2. Billy and Suzy

The route $\langle ST, SH_t, BSt, BSt' \rangle$ is active if we hold fixed BH_t' at its actual value ($BH_t'=0$). If $ST=1$, then $SH_t=1, BSt=1$ and $BSt'=1$. But if $ST=0$, then $SH_t=0, BSt=0$, and given that $BH_t=0, BSt'=0$. Hence BSt' counterfactually depends on the value of ST , holding fixed $BH_t'=0$.

$\langle BT, BH_t', BSt' \rangle$, on the other hand, is *not* active. Holding fixed that the bottle shatters at time t ($BSt=1$), Billy's Throw makes no difference to BSt' . If $BT=1, BSt'=1$; if $BT=0, BSt'=1$. Hence BSt' does not depend counterfactually on BT .

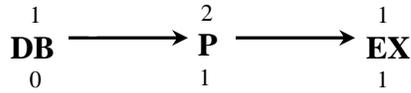
Hence Suzy's throw (but not Billy's) comes out as the cause of the bottle's being shattered at time t' , as we desire.

(H1) is thus able to successfully handle early and late preemption. What about Dog Bite and Boulder, which (L) ruled in as cases of causation as a result of adding the Transitivity condition?

2.7.2 Dog Bite

The causal model for Dog Bite is shown below in figure 2.3. If the dog bites, $DB=1$. If Terrorist pushes with his right hand, $P=1$; if with his left, $P=2$. If the Bomb explodes, $EX=1$. Now, if

DB=1, EX=1. And if DB=0, EX=1. Hence the explosion does not depend on the dog's biting of Terrorist's right hand, according to (H1), as we desire.



$$\begin{aligned}
 DB &= 1 \\
 P &= DB+1 \\
 EX &= P \text{ if } P \neq 2 \text{ and } 1 \text{ if } P=2
 \end{aligned}$$

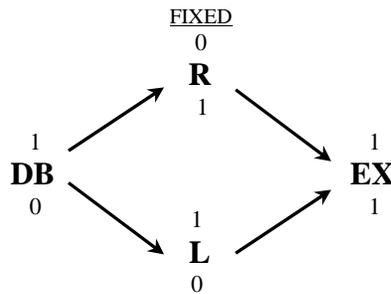
Values of P: 0,1,2 (Not pushed, right, left)

Is DB a cause of EX?

(H1) = No
Intuition = No

Figure 2.3. Dog Bite (3-variable)

If, however, we were to construct a four-variable model of Dog Bite (figure 2.4), this theoretical verdict would be reversed, the Holding Fixed move revealing a latent dependency between EX and DB. Let the binary variable R represent the Terrorist's pushing the button with his right hand, and L the left.



$$\begin{aligned}
 DB &= 1, L=DB \\
 R &= \sim DB \\
 EX &= L \vee R
 \end{aligned}$$

Is DB a cause of EX?

(H1) = Yes
Intuition = No

Figure 2.4. Dog Bite (4-variable)

If we now hold R fixed at its actual value ($R=0$), EX *will* depend counterfactually upon DB. For if $DB=1$, $L=1$ and hence $EX=1$. But if $DB=0$, $L=0$, and given that $R=0$ too, $EX=0$.

Hitchcock's reply is to claim that the four-variable model is *inappropriate*, and that the example is better represented by the three-variable model. He writes:

What makes [the four-variable representation] an inappropriate model for Dog Bite? One minimum criterion of adequacy for a model is that it entail only true counterfactuals. As it turns out, the very counterfactual that reveals the active route $\langle D, L, E \rangle$ in this model is false, or at best indeterminate. That...counterfactual is: given that Terrorist did not push the detonator button with his right hand, if the dog had not bitten Terrorist's right hand, the (he would not have pushed the button with his left hand either) the bomb would not have exploded. Given that he did not push the button with his right hand, why should the dog bite make any difference to whether he pushes it with the left? He wanted the bomb to explode; would he not push the button with his left hand regardless of whether the dog bit his right?...Terrorist's pushing the button with his left hand ceases to depend counterfactually upon the dog bite when we specify whether or not he pushed the button with his right hand (2001, p.292-3).

2.7.3 Boulder

(H1) appears to be able to deliver the intuitively correct verdict for our example of a Self-Canceling Threat, Boulder.

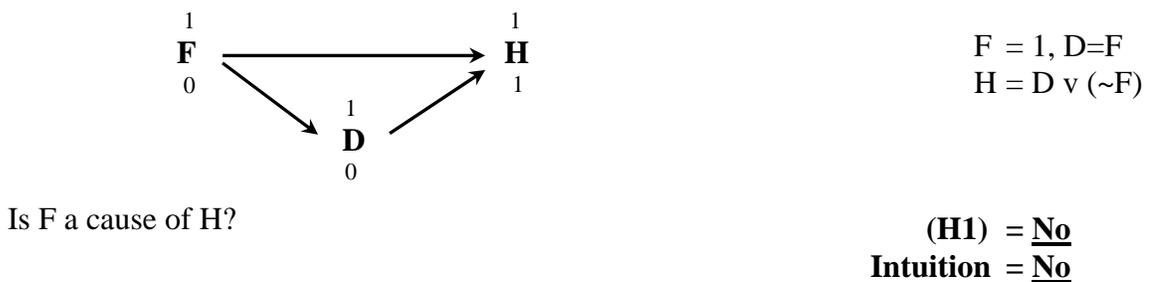


Figure 2.5. Boulder (3-variable)

Let F be the binary variable representing whether or not the boulder falls, let $D = 1$ if Hiker ducks, and $H=1$ if he survives. Holding fixed $D=1$, if $F=1$, $H=1$. And if $F=0$, $H=0$. So Hiker's survival does not depend on the boulder's falling, as there is no active route between F and H (figure 2.5).

But if we represent Boulder using a four-variable model, as Hitchcock notes, the theoretical verdict of (H1) is again reversed. Let the variable $A=1$ if the boulder reaches a point one meter from Hiker's head, so that even if he sees the boulder, it will be too late to duck. If we now hold fixed $A=1$, then H will depend on F . For if $F=1$, $D=1$ and so $H=1$. But if $F=0$, $D=0$ and so $H=0$. In words, holding fixed that the fact that the boulder is only a meter above Hiker's head, if the boulder hadn't fallen (and hence Hiker wouldn't have seen it in time to duck), Hiker wouldn't have ducked, and he would not have survived. Hiker's survival would then depend counterfactually on the boulder's falling (holding A fixed) and hence (H1) would rule that the boulder's falling caused Hiker to survive, which is strongly counterintuitive.

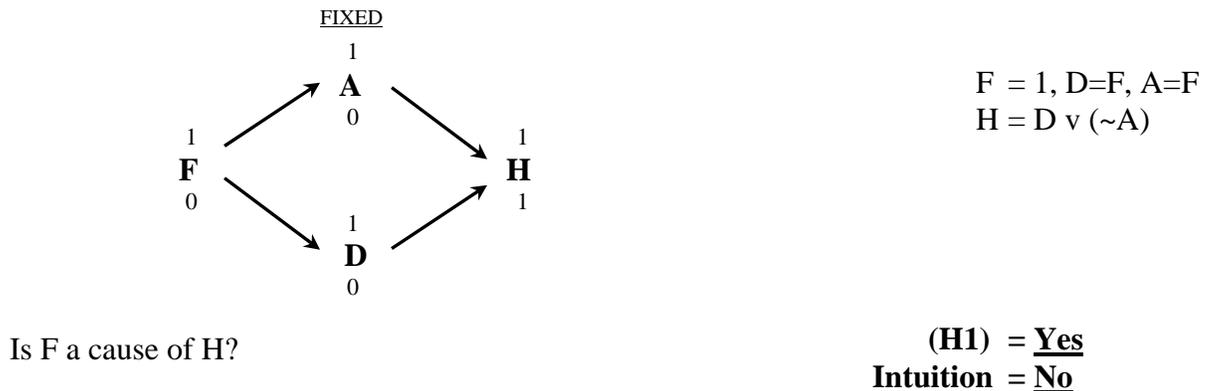


Figure 2.6. Boulder (4-variable)

Hitchcock again replies that the four-variable model is inappropriate, and that the variable A should be excluded from the model:

The interpolated variable A is not easy to find, and the...counterfactual that reveals the active causal route from F to H is not at all intuitive...[The] relevant piece of counterfactual reasoning...is correct, but bizarre. If the boulder never fell, how did it get to be there, one meter from Hiker's head? We are to imagine, presumably, that the boulder was mysteriously transported to a position immediately in front of Hiker's head. This is the sort of counterfactual reasoning that only trained philosophers engage in; unaided intuition is not to be faulted for failing to "see" the relevant...counterfactual. [W]e are not willing to take seriously the possibility that the boulder...comes to be in that position *even though the boulder does not fall in the first place*. This possibility is just too far-fetched (2001, p.297-8).⁶

He adds, in a related paper:

Once we introduce a variable...we admit the possibility that these variables might take on values independently of one another. When we choose to exclude a variable from a causal representation, it is because we do not take seriously the possibility that the variable could vary independently of the other variables. It is precisely for the purposes of contemplating hypothetical independent variations in the values of variables that they are explicitly included in a causal model (MSa).

Appropriate Causal Models

Let us attempt to summarize Hitchcock's criteria for deeming a causal model to be appropriate. In the context of a four-variable model with structural equations, as given in figure 2.7 below:

- (A1) The interpolated variable G should be easy to see so that it can be used in making a causal judgment.
- (A2) The model should entail no false counterfactuals. D must *not* cease to depend upon C when we fix G.

⁶ Some minor notational changes.

- (A3) It must not be considered too far-fetched for G to take its actual value when C takes its non-actual value. ‘G=1 when C=0’ must not seem too far-fetched.
- (A4) We admit the possibility that variables might take on values independently of one another. (In particular that G can vary independently of C).

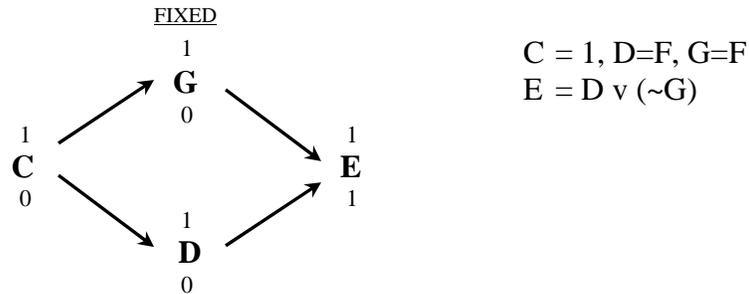


Figure 2.7. Model Appropriateness Criteria

Bearing these criteria of appropriateness (A1)-(A4) in mind, I will now examine some candidate counterexamples to (H1).

2.8 COUNTEREXAMPLES TO THE SUFFICIENCY OF (H1): SWITCHING AND SELF-CANCELING THREATS

2.8.1 First Counterexample to the Sufficiency of (H1): Two Trolleys

Consider the following case of ‘Switching’:

Flip: A trolley is hurtling down a track towards Victim. The track diverges into two 100-yard subtracks, which then reconverge. Victim is strapped to the track just beyond the reconvergence point. Just as the trolley is approaching the divergence point, Suzy flips a switch (F) that takes the trolley onto the left subtrack (L). The trolley travels the 100 yards of this subtrack, and then regains the main track, crushing Victim (V). Had Suzy not

flipped, the trolley would have taken the right subtrack (R), but Victim would have nonetheless been crushed.⁷

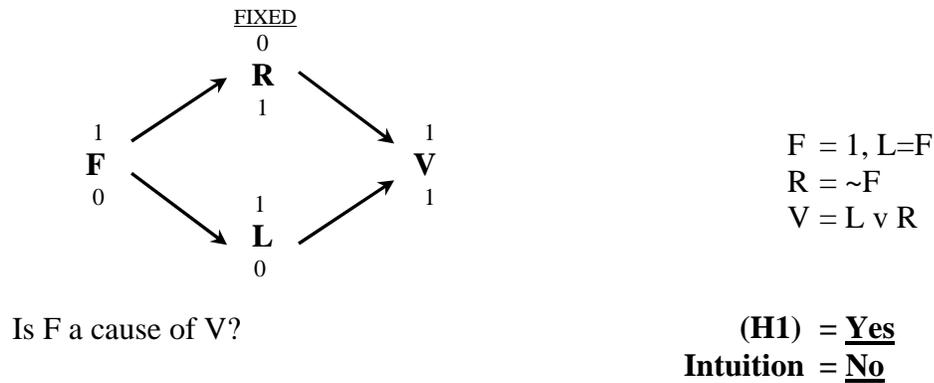


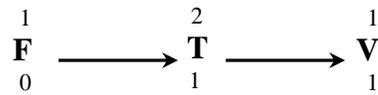
Figure 2.8. Flip

While it is intuitively obvious that the flipping is not a cause of Victim’s crushing, according to (H1), the flipping *is* a cause of the crushing: the crushing depends on the flipping, given that we hold fixed the fact that the trolley does *not* go down the right track. For if Suzy had not flipped ($F=0$), the trolley would not have gone down the left track ($L=0$) and as we are holding fixed the fact that the trolley did not travel down the right track ($R=0$), the trolley would have gone down neither. As these are the only two options the trolley has for reaching Victim, Victim would not have been crushed ($V=0$). Hence Flip counts as a *prima facie* counterexample to (H1).

Hitchcock has suggested that Flip is structurally equivalent to Dog Bite, and that therefore a four-variable model is inappropriate and should be replaced by a three variable model:⁸

⁷ Adapted from Yablo (2002).

⁸ Personal communication, 2005.



$$\begin{aligned} F &= 1 \\ T &= F+1 \\ V &= T \text{ if } P \neq 2 \text{ and } 1 \text{ if } T=2 \end{aligned}$$

Values of T: 0,1,2 (No track, right track, left track)

Is F a cause of V?

(H1) = No
Intuition = No

Figure 2.9. Flip (3-variable)

This three-variable model now delivers the correct verdict that the flipping is not of cause of Victim's crushing. But *is* the four-variable model inappropriate according to criteria (A1)-(A4)? First, the interpolated variables L and R are not hard to see.

Second, it does not appear that the following counterfactual entailed by the four-variable model is false: 'L depends on F when R is fixed at its actual value of zero.' If the trolley doesn't go down the right track, it is necessary that the flipping occurs in order to take it down the left track. We must not backtrack and say that if R=0 then the lever must have been flipped therefore the trolley had to go down the left track. So the counterfactual appears to be true. The reasons for thinking the corresponding counterfactual in Dog Bite is false seem to have to do with contingent details of the case: the Terrorist's wanting the bomb to explode and therefore wanting to push the button with his left hand if he had not done so with his right. But there does not appear to be anything analogous in Flip.

Third, it does not seem to far-fetched for the trolley not to go down the right track even though the flipping did not occur. We can imagine that the train could have derailed for example (although admittedly, this possibility is not explicitly stated in the facts of the case).

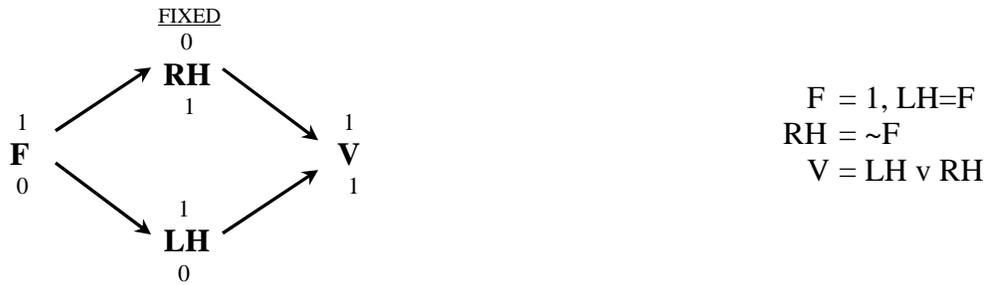
There does, however, appear to be something of a problem with criterion (A4), which is also linked to (A2). In including two variables (L and R), we should admit that they can take on values independently of one another. But one combination that is clearly not possible is for L and

R to both take the value '1'. In that case, the trolley would simultaneously be traveling down both subtracks! If we fix $R=1$, then this does appear to fix $L=0$. Perhaps this warrants opting for the three-variable model over the four-variable. Perhaps we should make explicit a fifth criterion of appropriateness:

- (A5) The model must not allow variables to take on values that are *logically incompatible*.

But consider now a slight variation on Flip:

Two Trolleys: Two parallel subtracks run alongside one another towards a movable section of track that is connected to a single continuing main track. The moveable section can be positioned so that it either connects with the right or left subtrack (it is initially connected to the right subtrack). A trolley is hurtling along each track towards the movable section. If a lever is flipped, the left subtrack will be connected to the main track, and the trolley that was traveling down the left subtrack will continue its journey along the main track. If the lever is not flipped, the left trolley will derail, but the trolley that was traveling down the right subtrack will continue onto the main track. Victim is strapped to the main track just beyond the flipping point. As the trolleys are approaching the flipping point, Suzy flips a switch (F) that takes the left trolley onto the main track; the right trolley derails. The left trolley hits Victim, who is crushed. Had Suzy not flipped, the right trolley would have continued onto the main track and would still have been crushed. Let us choose the following variables: RH (the right trolley's hitting victim), LH (the left trolley's hitting victim) and V (Victim's crushing).



Is F a cause of V?

(H1) = Yes
 Intuition = No

Figure 2.10. Two Trolleys

In this example, there is no logical incompatibility in assigning $RH=1$ and $LH=1$. It is logically possible for both trolleys to hit Victim. Nor does it seem too far-fetched to imagine either trolley derailing; this possibility is explicitly mentioned in the description of the example. Two Trolleys therefore poses a more serious challenge to (H1).

2.8.2 Second Counterexample to the Sufficiency of (H1): Two Assassins

Two Assassins: Captain and Assistant are on a mission to kill Victim. On spotting Victim, Captain yells “Fire!” and Assistant shoots at Victim. Victim overhears the order, and although the bullet almost hits him, ducks just in time and survives unharmed. Later, at a prearranged medical check-up, Victim receives a glowing health report. If Captain hadn’t yelled “Fire!”, Assistant would not have shot, and Victim would have received the same glowing report. If Victim had not ducked, however, he would have been killed, and would not have received the glowing report.⁹

⁹ Adapted from Hitchcock (2003) with modifications due to Yablo (2004). Example originally due to McDermott.

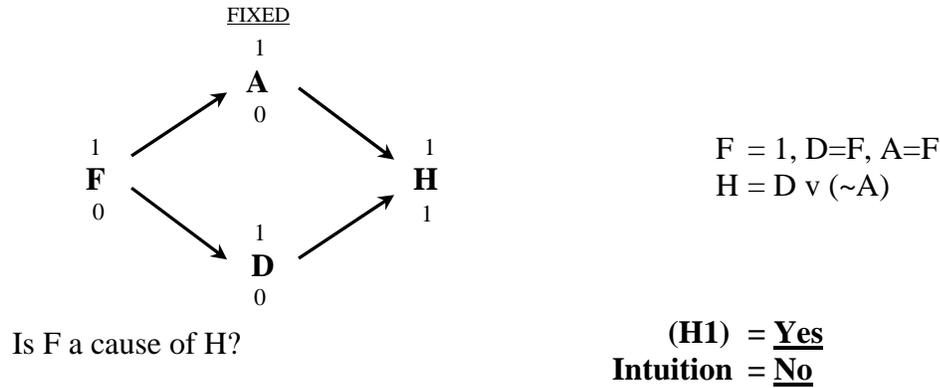


Figure 2.11. Two Assassins

Obviously, Victim’s receiving a glowing report does not depend *simpliciter* on Captain’s yelling “Fire!”; if Captain hadn’t yelled, there would have been no threat to Victim’s health, and he would have received the glowing report. But holding fixed that Assistant shoots, Victim’s receiving the glowing report not only depends on his ducking, but also, surprisingly, on the Captain’s yelling “Fire!”. For if Captain hadn’t yelled “Fire!”, Victim would not have heard the order. As a result, Victim would not have ducked. Given that Assistant shot (we are holding this fact fixed), Victim would have been killed and would not have received the glowing report. That the receipt of the report (H=1) depends on Captain’s yelling “Fire!” (F=1), holding fixed that Assistant shoots (A=1), can be seen from the causal model: if F=1 and A=1, then R=1; if F=0 and A=1, then R=0. In consequence, (H1) would judge Captain’s yell to be a cause of Victim’s receiving the report.

This case has the same counterfactual structure as the four-variable version of Boulder. Captain’s yelling “Fire!” is a self-canceling threat. Does the four-variable model of Two Assassins fall foul of the appropriateness criteria in the same way that Boulder does? In this case, the appropriateness conditions seem to be met. First, the variable A is not difficult to see, and second, the ducking does not cease to depend on Captain’s yelling when we fix whether or not Assistant shot. Third, it is not difficult to imagine Assistant firing even if Captain did not yell “Fire!”. Indeed, Lewis suggests that we grant human agents autonomy ‘by courtesy’.¹⁰ And we

¹⁰ Lewis (1986). Discussed in Hitchcock (MSa).

are willing to entertain the possibility of the four variables varying independently of one another. Two Assassins thus appears to be a *bona fide* counterexample to (H1).

Hitchcock's reply (on the basis of informal surveys) is that our intuitions are either divided or unclear with regard to this case.¹¹ My response is to flatly reject any intuition other than that which says that Captain's yelling "Fire!" is *not* a cause of Victim's survival (and subsequent receipt of an excellent health report). I argue that any inclination towards judging Captain's yelling "Fire!" to be a cause is simply mistaken. In my experience, those individuals who lean in this direction are brought around by reminding them of the counterfactual, 'If Captain hadn't yelled "Fire!", Assistant would *not* have shot.' On recalling this, individuals almost always immediately retract their initial intuition. It is not hard to understand why this mistake is sometimes made. The case is somewhat complex and esoteric, and difficult to manage cognitively. It is easy to *forget* that Assistant, *by stipulation*, would not have fired unless ordered. Second, we are generally *not certain* about what would have happened counterfactually, had the putative cause not occurred. Even though we are *told* in this case that Assistant would not have shot, there is a lingering feeling that he *might* have. It is thus understandable that Captain's yell comes to be seen as something of a warning against an *already present* danger. These factors, which distort our intuitions, are present to an even greater degree in the following structurally analogous case:

Avalanche Warning: A party of skiers have wandered into an unstable area where there is a very high risk of an avalanche. Fred, a Mountain Rescue officer who is patrolling the nearby area, notices the party and realizes the extreme danger they are in. He yells, "Get out of there!" and the party escape moments before an avalanche occurs.

It would seem clear that Fred's yelling was a cause of the party's survival. But let us stipulate that the avalanche would *not* have occurred had Fred not yelled; the avalanche is actually triggered by the sound waves of Fred's very loud yell. In this example, it is easy to see why one is tempted to view Fred's yelling "Get out of there!" as a cause of their survival. Fred's yell is *explicitly* a warning and the (epistemic) probability of the avalanche occurring had Fred not

¹¹ Hitchcock (2003, p.9-11)

yelled is very high.¹² Yet I maintain that the intuition that Fred was a cause is mistaken, and simply amounts to a refusal to accept the stipulated facts of the case. I submit that *no* Self-Canceling Threat should count as a cause.^{13, 14} Two Assassins thus stands as a counterexample to (H1).

The strategy of omitting variables from a causal model that would reveal troublesome latent dependencies between the *non*-cause and effect seems to be contrary to a fundamental motivation behind the Holding Fixed strategy, which *is* to make explicit dependencies that are hidden by the influence of other variables; this is why it works so well for preemption. I consider the selective suppression of variables from causal models to be a somewhat *ad hoc* maneuver that should be resisted. What *would* be welcome is some means of *restricting the domain of applicability* of the Holding Fixed strategy to only cases of Preemption, so that Switches and Self-Canceling Threats would *not* be counted as causes. One quite promising recent idea that would limit the use of Holding Fixed is due to Hitchcock himself (MSa, MSb). He employs what he dubs ‘The Principle of Sufficient Reason’ as a criterion for the selective application of Holding Fixed.

2.9 COUNTERFACTUAL DEPENDENCE AND THE PRINCIPLE OF SUFFICIENT REASON (H3)¹⁵

Hitchcock begins his discussion of the Principle of Sufficient Reason and its application to the counterfactual analysis of token causation, with quotations from Leibniz (unsurprisingly) and from Mill:

Nothing happens without a sufficient reason, why it should be so, rather than otherwise (Leibniz 1716, §2).

¹² Since we are assuming determinism, the probability of the avalanche occurring was zero.

¹³ Yablo (2004) also argues strongly in support of this view.

¹⁴ It is possible, in one sense, to describe preempting causes as Self-Canceling Threats. Trainee’s shot ‘threatens’ Victim’s bleeding to death by preventing Supervisor’s shooting. But Trainee’s shot also ‘cancels’ this threat by causing Victim’s death directly. The counterfactual structure of preemption cases is very different from the cases I am calling Self-Canceling Threats, however, and no confusion over the use of the term should occur.

¹⁵ I use (H3) rather than (H2) since (H2) has been used in the literature (e.g. Hiddleston 2005) to refer to a slightly weakened version constructed to accommodate symmetric overdetermination. I will not discuss (H2) here.

From nothing, from a mere negation, no consequences can proceed. All effects are connected, by the law of causation, with some set of positive conditions...(Mill 1843, Vol. I, Chapter V, §3).

The central idea of the proposal is that the relative locations of omissions and positive events (commissions) within a ‘causal network’ will serve as a criterion for when to apply, and when not to apply the Holding Fixed strategy. Hitchcock writes:

Perhaps our judgments of token causation are based not merely upon the form of the token causal structure, but also upon the location of positive events and omissions within that structure. I am skeptical that a rigorous distinction between positive events and omissions can be sustained...I prefer a slightly different distinction, between what I will call *default* and *deviant* outcomes. A default outcome, roughly, is one that we expect, or take for granted in a certain context. The distinction between default and deviant is pragmatic...(MSa).

Roughly, deviant values correspond to positive events, and defaults correspond to absences. I do not think that this distinction can be made in any precise way but will offer some rules of thumb:¹⁶

- (D1) The default value of a variable is the one that we would expect in the absence of any information about intervening causes.
- (D2) In particular, changes in persisting states are typically considered deviant outcomes, while the continuing persistence of a state is the default.
- (D3) Temporary actions or events tend to be regarded as deviant outcomes.
- (D4) Those outcomes that involve the absence of an entity are often regarded as defaults.
- (D5) In the case of human actions, we tend to think of those states requiring bodily motion as deviants, and those compatible with lack of motion as defaults.

¹⁶ Quoted with minor changes in structure. Numbering my own.

(D6) We typically feel that deviant outcomes are in need of explanation, whereas default outcomes are not necessarily in need of explanation.

In most cases, however, the assignment of default and deviant values is fairly natural. (Hitchcock, MSb)

The intuitive idea behind the Principle of Sufficient Reason (PSR) is that default values of causes cannot give rise to deviant values of their effects:

[(PSR) is satisfied iff]...If every parent¹⁷ of a variable X in a causal model $\langle V, E \rangle$ takes a default value, then X takes a default value (Hitchcock, MSb).

The following definitions are then offered:

(CN) Let $\langle V, E \rangle$ be a causal model, and let $X, Y \in V$. The *causal network* connecting X to Y in $\langle V, E \rangle$ is the set $N \subseteq V$ that contains exactly X, Y , and all variables Z lying on a path from X to Y in $\langle V, E \rangle$.

(SCN) A causal network is *self-contained* iff for all Z in N , Z takes a default value when all of its parents in N do.

That is, N is self-contained if every variable in N satisfies a restricted version of PSR, where only the parents that are themselves in N are relevant. Intuitively, a network is self-contained when it is never necessary to leave or augment the network to explain why the variables within the network take the values they do. (MSb).

Hitchcock then presents a revised account of token causation:

(H3) Let $\langle V, E \rangle$ be a causal model, let $X \in V$, and let $X=x$ and $Y=y$. If the causal network connecting X to Y in $\langle V, E \rangle$ is self-contained, then $X=x$ is a token cause of $Y=y$ in $\langle V, E \rangle$ iff Y counterfactually depends upon X in $\langle V, E \rangle$. (MSb).

¹⁷ Any variable which appears on the right hand side of the structural equation for X .

(H3) is not a full analysis, since it only applies to self-contained networks. Elsewhere, however, Hitchcock adds:

...[I]f Y fails to depend on X in a model that is not self-contained, the failure of counterfactual dependence may be due to the influence of extraneous factors. At this point, it is appropriate to look for active causal paths by holding fixed the variable which violates PSR. (This effectively negates the disturbing influence of the extraneous factors). (Hitchcock, MSa).

Let us, for the moment, put forward a strong version of (H3), which claims, in short, that:

1. In a self-contained network, C is a cause of E iff E depends on C.
2. In a non-self-contained network, C is a cause of E iff E depends on C with some PSR-violating variable held fixed at its actual value.

It is true that Hitchcock does not put (H3) forward in such a strong form as this; he leaves it open whether or not (H3) may be overridden by other considerations in certain circumstances. As we shall see towards the end of this chapter, Hitchcock thinks that (H3) may work *in tandem* with (H1) in determining our intuitive causal judgments. But first let us examine how this strong version of (H3) fares with the Preemption Problem and with Self-Canceling Threats.

2.10 (H3), PREEMPTION AND SELF-CANCELING THREATS

In our standard example of early preemption, Trainee and Supervisor, following (D1)-(D6) above, the default values of T, S and V are all zero. Note that when T takes a default value, S takes a deviant value (figure 2.12). S therefore violates (PSR) and the causal network (T,S,V) is *not* self-contained. Hence it is appropriate to hold it fixed and look for counterfactual dependency between V and T. Holding fixed S at its actual value of zero, V does depend on T, as with (H1). Hence (H3) delivers the right verdict for this case of early preemption. Intuitively, S

cannot take the deviant value ‘1’ for no (sufficient) reason. This indicates, Hitchcock claims, that there is some influence from outside the network – perhaps the orders that S received before embarking with T on their mission to kill Victim. Holding S fixed at its actual value screens off this external influence.

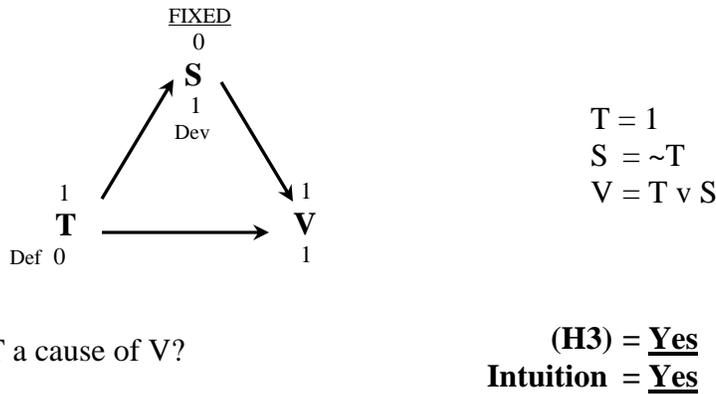


Figure 2.12. Trainee and Supervisor

(H3) is also able to handle our late preemption case, Billy and Suzy (figure 2.13).¹⁸ The default values of the variables BT, ST and BSt are all zero. There is a minor complication concerning the default value of the variable “Bottle Shattered at time t’ (BS t’).” The default value of BS t’ depends on the value of the BT. If the bottle is already shattered at time t, then our default expectation is that it will be shattered at t’. If the bottle is not shattered at t, our default expectation is that the bottle will be unshattered at t’. So the default value of BS t’ is equal to the value of BSt, i.e. Def (BS t’) = 1 when BSt = 1, and Def (BS t’) = 0 when BSt = 0.

In the causal network (ST, BSt, BS t’), BS t’ takes the value ‘1’ when its parent in the network, BSt, takes a default value, zero. When BSt takes the value zero, the default value of BS t’ is also zero. So BS t’ taking the value ‘1’ when its parent in the network, BSt, takes a default value,

¹⁸ I have omitted the variables SHt and BHt’ (cf. figure 2.2), as does Hitchcock in (MSb).

means that $BSt'=1$ is deviant. Hence BSt' violates (PSR). Holding fixed the violating variable at its actual value will *not* reveal the desired dependence between ST and BSt' . If however, we hold BT fixed at its non-actual value, '0', then the sought dependency can be made to appear. When $BT=0$, if $ST=1$, $BSt'=1$; if $ST=0$, $BSt'=0$. Hence Suzy's throwing is a correctly ruled to be a cause of the Bottle's being shattered at t' .

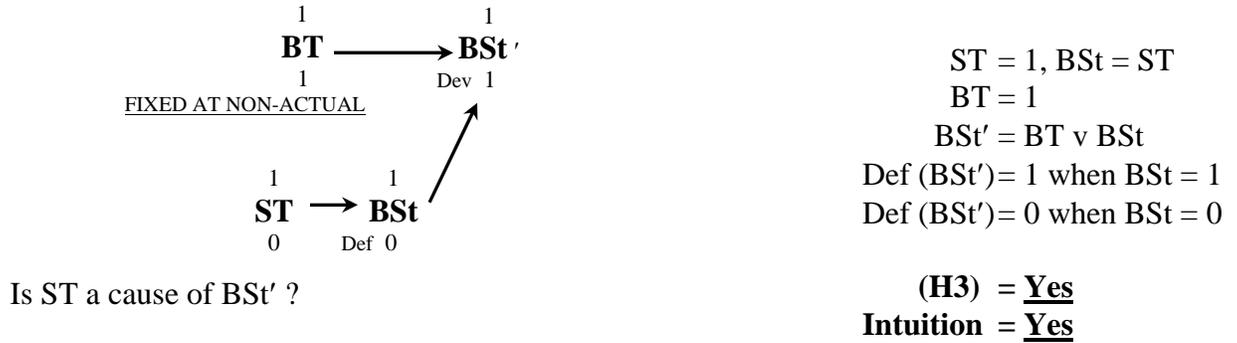


Figure 2.13. Billy and Suzy

This strategy is not entirely satisfactory; the disunity between the treatments of early and late preemption prompts one to think that (H3) requires further work. But let us put aside this worry for the moment and ask whether (H3) can deliver the right result for our Self-Canceling Threat example, Two Assassins.

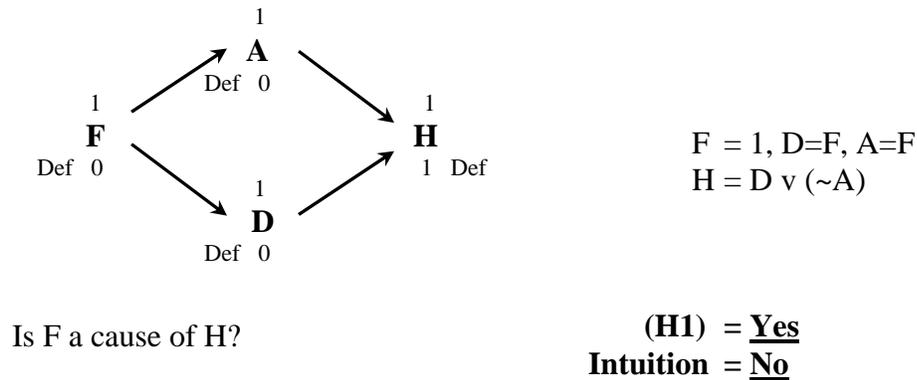


Figure 2.14. Two Assassins

The default values of all four variables, following (D1)-(D6), are all zero. No variable violates (PSR) since each one takes a default value when its parents do. Hence the network is self-contained, and, in contrast to our preemption examples, there is no reason to employ the Holding Fixed strategy in order to screen out dependency-breaking influences from outside the network. H does not depend simpliciter on F, so (H3) correctly rules that Captain's yelling "Fire!" is not a cause of Victim's subsequent good health.

So far, so good. However, as I will show in section 2.11, (H3) runs into trouble with cases of Switching. In addition, the introduction of (PSR) allows in *new* counterexamples involving even early preemption, as well as some Self-Canceling Threats.

2.11 COUNTEREXAMPLES TO (H3)

How might we go about generating counterexamples to (H3)? The most convincing counterexamples would present self-contained networks in which intuitively there *is* causation but *no* dependence, or, dependence but *no* causation. In non-self-contained networks, we could seek counterexamples where there *is* dependence, holding fixed some fact G, but which are intuitively *not* cases of causation.

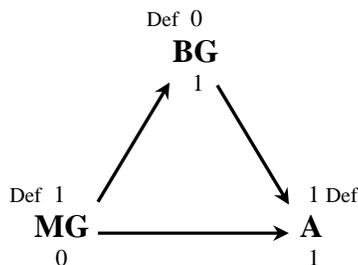
Let us begin with self-contained networks, and first try to find a case of causation without dependence. Preemption is the obvious starting point when looking for causation without dependence. Normally, in cases of preemption, the redundant backup violates (PSR) because it takes a *deviant* value when its parent (the preempting cause) takes a *default* value. In virtue of this violation, we hold the backup fixed, thereby revealing the latent dependency between the cause and the effect. If we could engineer a case in which the backup took a *default* value when the preempting cause took a default value, we would have a self-contained network, *without* dependence. (H3) would incorrectly judge this type of preemption *not* to be causation. One way generating such a case is to make the preempting cause's actual value correspond to its default value. How might a *default* value of a preempting cause bring about an effect? An example taken from Menzies (MS) seems promising:

2.11.1 First Counterexample to (H3) in a Self-Contained Model: Main Generators

Main Generators: A bank's alarm system (A) is powered by electricity from the city's main generators (MG). If the main generators were to fail, a small on-site backup generator (BG) would temporarily continue to power the alarm system at the bank. In fact, the main generators do *not* fail, and the alarm system remains on, as normal.¹⁹

Are the main generators a cause of the alarm system's being on? Intuitively it appears so. It is the electricity from the main generators (rather than the backup generator) that is supplying the power. This example just seems to be a straightforward case of preemption, with the main generators preempting the backup generator.

Let $MG=1$ if the main generators are on and let $MG=0$ if not. Let $BG=1$ if the backup generator is on and let $BG=0$ if not. And finally let $A=1$ if the alarm system is on $A=0$ if not. Our causal model is then:



$$\begin{aligned}
 MG &= 1 \\
 BG &= \sim MG \\
 A &= MG \vee BG
 \end{aligned}$$

Is MG a cause of A?

(H3) = No
 Intuition = Yes

Figure 2.15. Main Generators

The crucial feature of this example is that the main generators' being on, while clearly the cause of the alarm system's being on, is the *default* state of the system. That this is so can be seen, via (D1)-(D6), from the fact that main generators' being on:

1. Is the expected state of the system in absence of evidence to the contrary.

¹⁹ Adapted from Menzies, 2004 draft.

2. Involves persistence of that state (rather than change).
3. Is not a temporary action.
4. Does not particularly call for explanation.

The default value of the background generator, BG is clearly zero, whereas the default value of the alarm system A is clearly '1', for the same reasons that the default value of MG is '1'. When MG takes its default value of '1' (in the actual scenario), BG *also* takes its default value of zero. And when A's parents take their default values, so does A. Hence no variable in the causal network (MG, BG, A) violates (PSR). The network is therefore self-contained, and we thus have no reason to try to screen out any possible external influences on the network by holding any variable fixed. This is bad news for (H3), since in order to generate the correct theoretical verdict (that the main generators are the cause of the alarm system's being on), we must hold BG fixed at its actual value of zero, in order to reveal the dependence of A on MG. Since we cannot do this, Main Generators stands a clear counterexample to (H3).

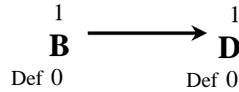
Although I will not do so here, it is a simple exercise to create a whole family of counterexamples similar to Main Generators. There are many causal 'agents' that are analogous to the main generators, whose causal efficacy stems from their *default* state. For example, radioactive materials decay 'by default' (according to (D1)-(D6)), the Sun shines by default, poison gases diffuse by default, the void is deadly by default, and so on. To construct counterexamples we simply need to make these agents preempt similar backups.

In addition, we can generate further counterexamples by swapping the positions of MG and BG in figure 2.15. Imagine that once a week, the bank carries out a brief test of its backup system. For ten minutes, the bank switches over to the backup generator, during which time, if it were to fail, the main generators would immediately come back on. In this case, we have another straightforward preemption, this time of the main generator by the backup. Now, when BG takes its default value of zero, MG takes its default value of '1' (in a scenario in which the backup generator fails). Again, no variable violates (PSR), and we again have a self-contained network without dependence, but which clearly constitutes a genuine case of causation. We thus have a further counterexample to (H3).

Next, let us attempt to find counterexample that contains a self-contained network in which there *is* dependence, but that is intuitively *not* a case of causation.

2.11.2 Second Counterexample to (H3) in a Self-Contained Model: Birth and Death

Birth and Death: Was Newton's birth (B) a cause of his death in 1727 (D)?



$$\begin{array}{l}
 \mathbf{B} = 1 \\
 \mathbf{D} = \mathbf{B}
 \end{array}$$

Is B a cause of D?

(H3) = Yes
Intuition = No

Figure 2.16. Birth and Death

The default values of the events B and D are zero, unproblematically. A when B takes a default value, so does D. Hence the network is self-contained. According to (H3), since D depends on B, Newton's birth was a cause of his death in 1727, which seems wildly counterintuitive. This example is problematic for *any* counterfactual analysis of causation, and similar cases have generated a fair amount of discussion in the literature. Lewis (2004), for example, claims that we just do not think to mention the birth as a cause (though strictly it is). One might also reply to the Birth and Death counterexample by suggesting that the birth doesn't count as a cause since it is *too widely separated in time and space*. Ned Hall (2004) has suggested that causation may come in *degrees*, and that with increasing distality of the putative cause from the effect, we reach a point where we no longer consider it to be a cause, even though the effect still depends on it. But distality alone is insufficient to rule out an event as a cause: take a series of dominos and make it as long as you like. Conceptually it is perfectly clear that the first domino's toppling is a cause of the subsequent fallings of all the other dominoes in the chain. It might be replied that this case is not analogous to Birth and Death since the events that mediate between birth and death are extremely heterogeneous, whereas the domino topplings are all of a type. But consider a 'Rube Goldberg' machine that is as complicated as you like, with any number of heterogeneous events mediating between the initial lever pulling (or what have you) and its subsequent distal effect. Again, we have no difficulty in identifying the lever pulling as a cause. Now this might be a

result of our conceptualizing this series of heterogeneous events as a single ‘machine’, in a way that we do *not* do for the set of events that mediate between life and death. I will not continue to discuss this case here, and the many possible approaches one might take towards defusing its threat, but leave it as a *prima facie* worry for the counterfactualist.

Consider next *non*-self-contained networks. To generate counterexamples, we need to find cases in which there where there *is* dependence (holding fixed some fact G), but which are intuitively *not* cases of causation. Switching and Self-Canceling Threats are the most obvious such cases. But as we saw with Two Assassins, its causal network was self-contained. Self-Canceling Threats would be problematic if we could construct an example in which the network was *not* self-contained. In Two Assassins, the variable ‘Assistant Shoots’ (A) does not violate (PSR), since it takes its default value of zero when Captain’s yelling “Fire!” takes *its* default value of zero. If we could engineer a case in which a variable analogous to A took a *deviant* value when its parent took a default value (and *vice versa*), then the variable A *would* violate (PSR). Holding A fixed would then reveal a latent counterfactual dependency, in virtue of which the example would be judged causal by (H3), and we would have a counterexample.

2.11.3 First Counterexample to (H3) in a Non-Self-Contained Model: HAL

HAL: Frank exits the Odyssey in his spacesuit in order to replace a faulty component on the ship’s exterior hull. During his spacewalk, Frank is temporarily fed oxygen through an oxygen line which extends from the ship through the airlock. The ship’s onboard computer HAL, who considers Frank to be a threat to the ship’s mission, closes the door to the airlock, severing Frank’s oxygen line. Dave, observing HAL’s actions, acts quickly and rescues Frank. Frank is later examined and found to be in good health.

Let HAL=1 if HAL cuts Frank’s airline and ‘0’ if not. Let A=1 if Frank receives the temporary oxygen supply and ‘0’ if not. Let D=1 if Dave rescues Frank and ‘0’ if not. And let H=1 if Frank receives a satisfactory health report and ‘0’ if not.

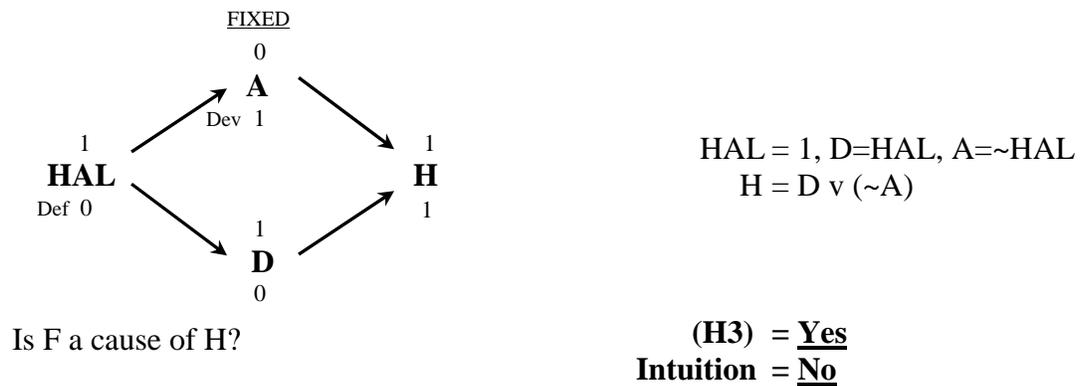


Figure 2.17. HAL

Are HAL's actions a cause of Frank's passing his medical exam? Absolutely not; HAL seems to be a familiar kind of Self-Canceling Threat. Yet (H3) delivers the opposite verdict. This example differs from Two Assassins in that the threat introduced by HAL is a 'default' threat. It is the threat that the deadly void has by default. When HAL takes a default value of zero, Frank receives a temporary supply of air ($A=1$). This state of affairs is highly deviant in the vacuum of space. Frank's receiving air in this hostile environment is *not expected* in the absence of information; it requires explanation (that he is receiving air through an oxygen line); it is a temporary rather than a persisting state. The absence of air is clearly the default state; its presence is deviant. Hence A violates the Principle of Sufficient Reason, and the causal network (HAL, A, D, H) is not self-contained. Holding A fixed reveals the latent dependency between Frank's passing his medical exam and HAL's cutting Frank's oxygen line, and (H3) therefore judges HAL's actions to be a cause of Frank's good health. HAL is thus a counterexample to (H3). Note, for future reference, that HAL is also a clear counterexample to (H1). The appropriateness criteria (A1)-(A5) are satisfied. For example, it is easy to imagine Frank's air supply ceasing independently of HAL's actions.

Again, we could construct a family of related counterexamples in which the introduced threat threatens in its default mode. Several threats of this nature come to mind: radioactive decay, diffusion of poison gas, and so on.

Our final counterexample is our familiar case of Switching, Two Trolleys.

2.11.4 Second Counterexample to (H3) in a Non-Self-Contained Model: Two Trolleys

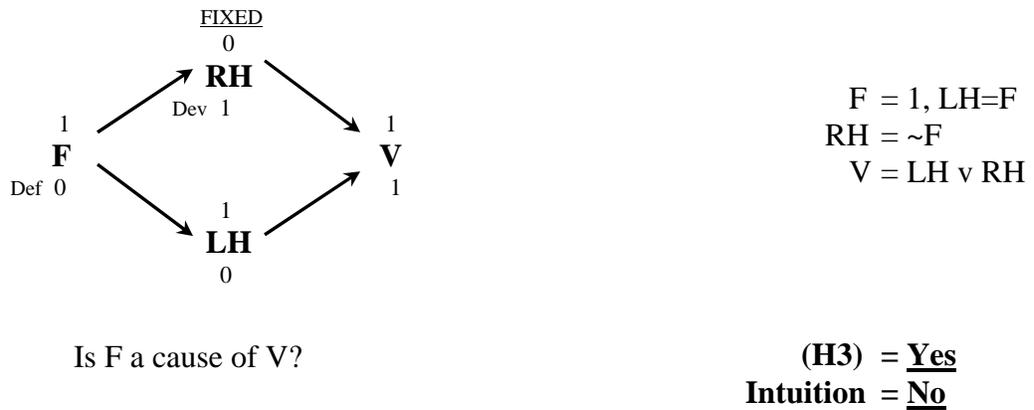


Figure 2.18. Two Trolleys

When Suzy's flipping takes a default value (i.e. she doesn't flip, and $F=0$), the right trolley hits Victim ($RH=1$). The default value of RH , however, is obviously zero, and therefore $RH=1$ is deviant. Hence RH violates PSR, and holding it fixed reveals the latent dependency between V and F . (H3) therefore delivers the incorrect verdict that Suzy's flipping is a cause of Victim's crushing.

The general principle that underlies the generation of these counterexamples is that (H3) will not work in causal networks where the distribution of deviant and default values is atypical. In such networks, (H3)'s hope of using the distribution of default and deviant values as a method for distinguishing between: (1) External influences to the network that we want to hold fixed (as in preemption), and (2): External influences that we do *not* want to hold fixed (in Switching and Self-Canceling Threats). The counterfactualist's hope of solving the Preemption Problem without incurring any penalties has been dashed again.

Table 1 below summarizes the verdicts of (H1) and (H3), together with our intuitive judgments (I) for the cases discussed in this chapter. Asterisks indicate theoretical verdicts that clash with intuition. Both (H1) and (H3) are found wanting.

Table 1. Theoretical Verdicts of (H1) and (H3)

| Case | <u>Trainee and Supervisor</u> | <u>Billy and Suzy</u> | <u>Two Assassins</u> | <u>Main Generators</u> | <u>Birth and Death</u> | <u>Two Trolleys</u> | <u>HAL</u> |
|-------------|-------------------------------|-----------------------|-----------------------|------------------------|------------------------|---------------------|-----------------------|
| Type | Early Preemption | Late Preemption | Self-Canceling Threat | Early Preemption | Distal Dependence | Switching | Self-Canceling Threat |
| (H1) | Y | Y | Y* | Y | Y* | Y* | Y* |
| (H3) | Y | Y | N | N* | Y* | Y* | Y* |
| I | Y | Y | N | Y | N | N | N |

2.12 (H3) AND (H1) AS MUTUALLY-REINFORCING INFLUENCES ON CAUSAL JUDGMENTS

To be fair to Hitchcock, (H3) was not put forward as a *full analysis* of token causation. Instead, he suggests that (H3) plays *some* role in influencing our intuitive causal judgments: (H3)'s verdicts can sometimes be overridden by other factors, he claims; in a similar fashion, (H1) may be viewed not as a complete analysis, but as another criterion that may influence our judgments:

Since I am not attempting to give an analysis of token causation, there is no need to specify a Boolean combination of the two rules [(H1) and (H3)] that yield necessary and sufficient conditions for causation. Both rules play some role in influencing our decision to call one event a token cause of the other. When the two rules agree, their verdicts reinforce each other, and our judgments are particularly clear. When the two rules disagree, they engage in a kind of destructive interference. The result is that our judgments become less clear, and more sensitive to other pragmatic considerations (Hitchcock, MSb).

I do not think that even this much-weakened position is defensible. As the table above indicates, (H1) and (H3) *both* deliver the wrong verdicts for HAL, Two Trolleys and Birth and Death. And when (H3) disagrees with (H1) in the Main Generators case, our judgments are not at all unclear.

2.13 CONCLUSION

(H1), while it is successful in dealing with the Preemption Problem, faces a variety of counterexamples. The most recalcitrant of these involve Switching and Self-Canceling Threats. With the introduction of the Principle of Sufficient Reason, (H3) fares somewhat better with these cases, but generates new counterexamples that (H1) dealt with easily, including simple cases of early preemption.

In the next chapter, I discuss another impressive technical attempt to solve the Preemption Problem within a counterfactual framework, using the Holding Fixed strategy. Yablo's *De Facto Dependence* theory of causation (2002, 2004) uses the notion of *artificiality* to handle difficult cases such as Switching and Self-Canceling Threats. The principal idea of his account is that some of the hidden dependencies that are revealed by holding fixed some fact G have an "artificial" quality, in virtue of which they should *not* be taken to indicate causation. The hope is that the notion of artificiality will be able to distinguish between the latent dependencies in preemptions and those in Switching and Self-Canceling Threats, treating the former as genuinely causal, and the latter as artificial.

3.0 DE FACTO DEPENDENCE

3.1 INTRODUCTION AND OVERVIEW

In this chapter I present several counterexamples to Yablo's *De Facto* Dependence theory of causation (DF) (2002, 2004). Like Hitchcock's active route theory, the major theoretical move is an attempt to formulate a more sophisticated version of the Holding Fixed approach. Self-Canceling Threats and Switching, as we saw in chapter two, are especially problematic for Hitchcock's theories. Yablo's introduction of the notion of *artificiality* offers hope of ruling out Self-Canceling Threats and Switching as *bona fide* forms of causation, while still delivering the intuitively correct verdicts for early and late preemption.

In section 3.2, I provide a brief outline of the De Facto Dependence theory. In section 3.3 I develop a new heuristic notation which enables (DF) to be represented in a causal modeling framework. In sections 3.4, 3.5 and 3.6, I show that (DF) appears to deliver the right results for familiar cases of Early and Late Preemption, Switching and Self-Canceling Threats. In section 3.7, I offer several counterexamples to the *sufficiency* of (DF), suggest a possible repair, hinted at by Yablo himself, but show that the repaired account cannot deal successfully with a variety of uncontroversial cases. In section 3.8, I propose a number of counterexamples to the *necessity* of (DF), discuss some possible replies, and show that (with the exception of a successful treatment of multiple redundancy), these replies are inadequate. As has been the case with all counterfactual theories of causation, the greatest challenge to (DF) comes from various kinds of preemption. In section 3.9, I show that there are examples for which (DF) is *undecidable*; that is, cases in which the theory does not deliver a determinate verdict on whether or not some putative C causes E, but for which our intuitions are clear. Finally, in section 3.10 I offer some concluding remarks about the future prospects of counterfactual theories in providing a univocal analysis of token causation.

3.2 DE FACTO DEPENDENCE

Yablo proposes the following definition of causation:

(DF) C is a cause of E if and only if E *de facto* depends on C,²⁰

where E *de facto* depends on C if and only if:

- (1) G puts E “in need of” C (i.e. E depends on C holding fixed some *fact* G)
- (2) There is no fact H that is *more natural* than G that makes E’s need for C *artificial*.

Condition (2) requires some explication. E’s need for C is artificial iff C neither *meets nor cancels any fallback needs*. What is a fallback need? Assume that history has a branching time structure, and that one branch, the *actual scenario*, corresponds to the way things actually developed after the occurrence of C. Another branch, which diverges from actuality, corresponds to what would have developed had C not occurred. What would have happened (counterfactually) after the branch point is the *fallback scenario*. A *fallback need* is an *event* that the effect E would have depended on in the fallback scenario. E’s *actual needs* are the events it depends on in the actual scenario, given that we hold G fixed. Note that the actual and fallback needs are limited to events that occur *after* the branch point. In Trainee and Supervisor (see chapter two for the details of this case), the fallback needs (for Victim’s bleeding to death) are Supervisor’s shooting and his bullet’s striking Victim. We could also include other events as needs, such as the bullet’s piercing Victim’s flesh, and so on. The actual needs are for Trainee’s shooting, and the corresponding events leading up to Victim’s bleeding to death.

Paraphrasing Yablo (2002. p.136-7), C is said to *meet* a fallback need *f* if C is a *counterpart* of *f*. This notion of a counterpart is best understood via examples. Trainee’s shot is a counterpart of Supervisor’s shot. Similarly, Trainee’s bullet’s striking Victim is a counterpart of Supervisor’s bullet’s striking Victim. If actual events can be matched up with fallback events “in such a way that salient features of the case are preserved (e.g. energy expended, distance

²⁰Yablo (2002) p.138

traveled, time taken, etc.)”,²¹ then these events are said to be counterparts. Trainee’s shooting is thus said to meet the fallback need for Supervisors shooting. If some of the fallback needs do not have counterpart needs in the actual scenario, those fallback needs are said to be *cancelled* by C. Consider the following Flip variant: Suzy flips the trolley over to a subtrack that is 100 yards long, where the untraveled subtrack is 150-yards long. Then the fallback need for the trolley to cover the extra 50 yards is cancelled by the flipping, because this need has no counterpart in the actual scenario.

If E depends on C in the actual scenario holding fixed some G (i.e. C is one of E’s actual needs), but C neither meets nor cancels any fallback need, E’s need for C is *artificial*. If we are unable to find a more natural (than G) fact H that puts E in need of C but does not make that need artificial, then condition (2) of (DF) is not met and C is not judged by (DF) to be a cause of E. In other words, there could be a fact G that makes E’s dependence on C artificial, but this does not automatically mean that C is not a cause of E. There may be another fact that we could hold fixed (H) that would reveal E’s dependence on C *not* to be artificial, in which case (DF) would rule that C *is* a cause of E. Yablo does not address the question of what it means for some fact G to be natural, and to what extent the notion is culturally and cognitively dependent. He holds out hope that it should in general be sufficiently clear whether G is more natural than H.²² In practice, however, the issue of naturalness rarely causes any great difficulty. In most cases, there is a rather obvious fact that will put E in need of C, and when this need is found to be artificial, one almost never finds a *more* natural G that also puts E in need of C but does *not* make the need artificial.

In simple cases of non-redundant causation, the effect does not even occur in the fallback scenario, and hence there can be no fallback needs, in which case there is no danger of artificiality. In such cases, E is put in need of G by the null fact ϕ . This need *cannot* be shown to be artificial, for there is no more natural fact H that might expose this need as artificial; no fact H is *more* natural than ϕ . Hence (DF) delivers the verdict that C is a cause of E in cases of non-redundant causation. For example, Fred throws a rock at a window and it breaks. The window’s breaking depends on Fred’s throwing holding fixed *nothing* (i.e. ϕ). The actual needs are for Fred’s throw, and this need cannot be shown to be artificial. This is the degenerate case of (DF).

²¹ Yablo (2002) p.137

²² Yablo takes the idea that naturalness is a respectable (and useful) metaphysical notion from Lewis’s “New Work for a Theory of Universals (1983).

In practice, the following (DF*) almost always delivers the same verdicts as (DF):

(DF*) C is a cause of E if and only if

(1*) E depends on C holding fixed some fact G.

(2*) C meets or cancels some fallback need.²³

For the majority of the putative counterexamples to (DF) presented in this chapter, I have constructed them in a manner such that while E does not depend on C *simpliciter*, it *does* depend on C, holding fixed some fairly *obvious* fact G. In testing (DF), we will therefore be mainly concerned with whether or not condition (2*) holds, i.e. whether C meets or cancels any fallback needs. The bulk of the evaluation of the putative counterexamples will consist of two steps: First, systematically checking whether C meets any fallback needs, and second, whether it cancels any. To summarize: In the examples presented in this paper, it is the case (with very rare exceptions) that if C meets or cancels some fallback needs, C is a cause of E. If C does neither, it is not a cause of E.²⁴

3.3 DE FACTO DEPENDENCE AND CAUSAL MODELS

It is sometimes difficult to keep track of the rather tortuous counterfactual reasoning involved in generating the theoretical verdict of (DF). The process is greatly facilitated by presenting (DF) in the causal modeling framework of equations and graphs presented in chapter two. I suggest the following additional notations in order to facilitate the generation of (DF)'s verdict:

1. Each event on which the effect depends (in either the actual or fallback scenario) is represented by a variable **V**. The value of the variable is 1 if the event occurs in the actual scenario (or would have occurred in the fallback scenario), and 0 if it did not occur in the

²³ (DF*) is only intended as a pedagogic and heuristic shorthand. It is not equivalent to (DF) but is certainly a lot easier to cognitively manage when testing it against candidate counterexamples.

²⁴ Again, this is not exactly equivalent to (DF). In cases where there is a disparity of any consequence between (DF) and (DF*), I will make this explicit and revert to (DF)

actual scenario, (or would not have in the fallback scenario).²⁵ The value a variable takes in the actual scenario (α) is written *above* the variable; the value it would have taken in the fallback scenario ('CF') is written below:

$$\begin{array}{r} \alpha \\ \mathbf{V} \\ \text{CF} \end{array} \begin{array}{l} 1 \\ \\ 1 \end{array} \begin{array}{l} \text{(Actual value)} \\ \\ \text{(Fallback value)} \end{array}$$

2. The structural equations show the dependent variable (on the left of the equality sign) as Boolean combinations of the variables on the right. In the causal graph, the variables are linked by directed arrows. "Negative" relationships are indicated with a '-' sign adjacent to the relevant arrow. If a variable is *exogenous* (i.e. has no directed edge running into it), it is given the value 0 or 1 in the structural equations in which it appears.

3. Counterparts are presented as ordered pairs of needs: actual need f first, fallback need g second: $\langle f, g \rangle$.

4. Needs are underlined and boldfaced.

If a fallback need is met by C, it is given the subscript 'm'; if cancelled, the subscript 'c':

$$\begin{array}{l} \underline{\mathbf{1}} \text{ Actual need} \\ \mathbf{V} \\ \underline{\mathbf{1}} \mathbf{m} \text{ Fallback need, met} \end{array} \quad \begin{array}{l} 1 \text{ Actual value} \\ \mathbf{V} \\ \underline{\mathbf{1}} \mathbf{c} \text{ Fallback need, cancelled} \end{array} \quad \begin{array}{l} 1 \text{ Actual value} \\ \mathbf{V} \\ \underline{\mathbf{1}} \text{ Fallback need, neither met nor cancelled} \end{array}$$

5. Actual and fallback needs are also listed in the following manner:

Fallback needs: $\mathbf{FAN} = \{X_m, Y_c, Z, \dots\}$ X met, Y cancelled, Z neither met nor cancelled.

Actual needs: $\mathbf{GAN} |G=\alpha = \{\dots, \dots\}$ Holding G fixed at its actual value.

6. Actual and fallback needs are also presented separately on decomposed or *partial* casual graphs, below each standard causal graph. Here the variables on which the effect depends

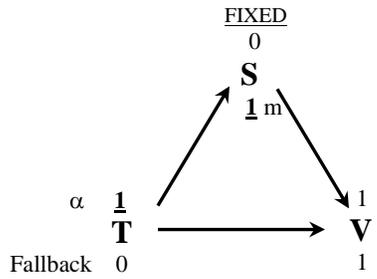
²⁵ Note that this characterization is for binary variables only.

(holding G fixed) in the actual scenario are presented together on one partial causal graph, and the variables on which the effect would have depended on in the fallback scenario (holding nothing fixed) are presented on another partial causal graph. For any event x , \mathbf{X} on the partial causal graph represents its occurrence and $\sim\mathbf{X}$ its non-occurrence. A fallback need that is cancelled is given the subscript 'c' as before; a fallback need that is met is given the subscript 'm'. A putative cause that neither meets nor cancels any fallback need is given the subscript 'xs' (short for excess) to indicate its artificiality.

Employing this notation, it can be seen immediately from the causal graph what the effect's needs are, and which (if any) are met or cancelled. For any graph in which *any* subscript appears, C is a cause of E. The partial causal graph pairs enable one to see at a glance, in an even more visually striking manner, whether or not actual needs exceed fallback needs. Presenting examples graphically in this way facilitates systematic checking of any pair of variables for possible dependencies and hence quickly determining the theoretical verdict of (DF); it is also a heuristically powerful representation that facilitates the generation of further candidate counterexamples.

3.4 EARLY AND LATE PREEMPTION

The annotated causal graphs for Trainee and Supervisor and Billy and Suzy (see chapter two for details: notation unchanged) are presented in figures 3.1 and 3.2 below.



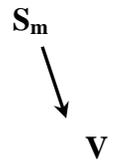
$T = 1$
 $S = \sim T$
 $V = T \vee S$
Counterparts : $\langle T, S \rangle$

FAN = $\{S_m\}$
GAN | $S=0$ = $\{T\}$

Is T a cause of V?

(DF) = Yes
Intuition = Yes

Fallback



Actual

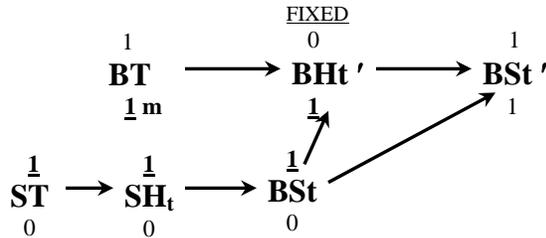


Figure 3.1. Trainee and Supervisor: Annotated Causal Graph

As with (H1), V depends on T holding fixed the fact that Supervisor did not shoot. But is Trainee’s shot artificial, and hence not a cause of V? No, because Trainee’s shot and Supervisor’s shot are counterparts, and thus Trainee’s shot *meets the fallback need* for Supervisor’s shot (indicated by the subscript ‘m’ in figure 3.1). In virtue of this, Trainee’s shot counts as a cause of Victim’s bleeding to death.

In Billy and Suzy, Suzy’s throw is a counterpart of Billy’s throw. Suzy’s throw meets the fallback need for Billy’s throw, in virtue of which Suzy’s throw is a cause of the bottle’s shattering. There is a small difficulty that needs to be cleared up, however. Yablo sometimes talks of putative causes being artificial when they “puff up the effect’s needs beyond necessity”

(2002, p.136). This is potentially confusing. The actual needs *do* exceed the fallback needs in this case, due to the presence of the extra variable BSt (as distinct from BSt'). In a sense the fallback needs *have* been “puffed up” by the putative cause, Suzy’s throw. But putative causes are *only* artificial if they neither meet nor cancel fallback needs. Suzy’s throw *does* meet a fallback need.



$$\begin{aligned}
 ST &= 1, SH_t = ST \\
 BSt &= SH_t \\
 BT = 1, BH_t' &= BT \& (\sim BSt) \\
 BSt' &= BH_t' \vee BSt
 \end{aligned}$$

Counterparts : <ST, BT>, <SH_t, BH_t>

$$\begin{aligned}
 \mathbf{FAN} &= \{BT_m, BH_t'\} \\
 \mathbf{GAN} \mid BH_t'=0 &= \{ST, SH_t, BSt\}
 \end{aligned}$$

Is ST a cause of the *state* BSt' ?

(DF) = Yes
Intuition = Yes

Fallback



Actual

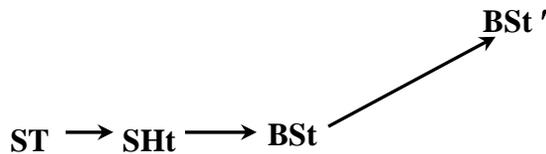
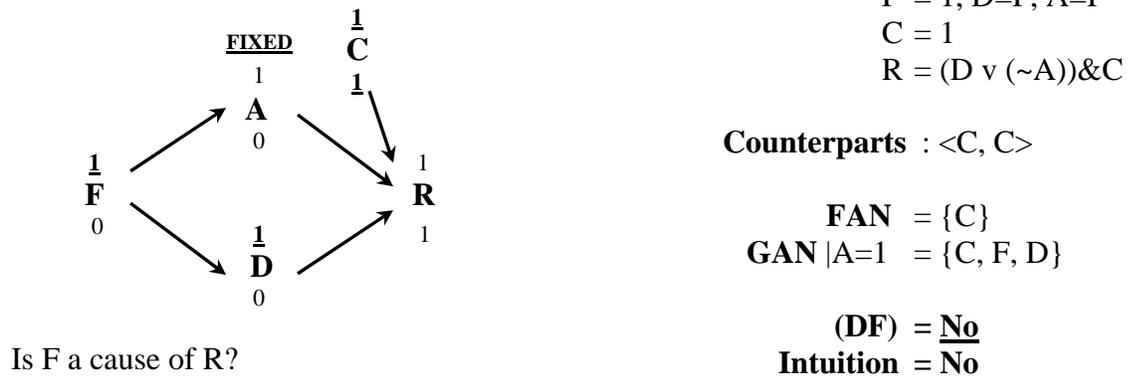


Figure 3.2. Billy and Suzy: Annotated Causal Graph

3.5 SELF-CANCELING THREATS

(DF) appears to get the right result for Two Assassins (again, see chapter two for the details of the case).²⁶ The fallback needs (FAN) are the events that the glowing report would have depended on, had Captain not yelled “Fire!”. We need not list them all, but let us focus on Victim’s doctor’s carrying out the health check (C). A counterpart of this need occurs, we assume, among the effect’s actual needs (the events the report depends on in the actual scenario). The actual needs (GAN), given that Assistant’s shooting (A=1) is held fixed, are: Captain’s yelling “Fire!” (F=1), Victim’s ducking (D=1) and the health check (C=1). These needs are represented on the annotated causal graph in Figure 5 below.



Fallback



Actual

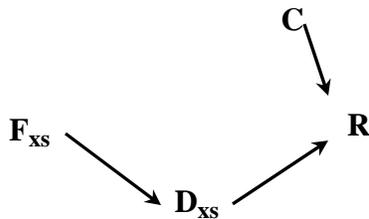


Figure 3.3. Two Assassins: Annotated Causal Graph

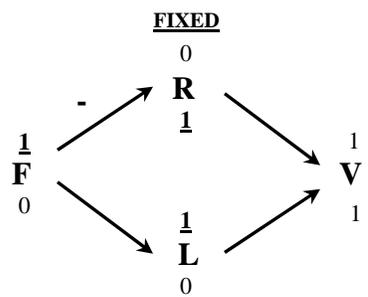
²⁶ In the next part of the paper, however, I show that this result involves a subtle mistake.

In virtue of the “recurrence” of the fallback need (C=1) in the actual scenario, Captain’s yell does not cancel any fallback needs. Neither is there a counterpart of the yell in the fallback scenario; hence the yell does not meet any fallback need. The yell therefore neither meets nor cancels any fallback needs, and the glowing report’s need for the yell in the actual scenario is therefore artificial, in that it is “piled on top” of the fallback needs. Is there any other, more natural fact than Assistant’s shooting that puts the glowing report in need of the yell, but which does not make this need artificial? None that come to mind readily. Hence condition (2) of (DF) is not satisfied, and Captain’s yelling “Fire!” is correctly judged by (DF) not to be a cause of Victim’s receiving the glowing report.

3.6 SWITCHING

(DF) also delivers the intuitively correct verdict for Flip. The fallback needs are for the trolley to successfully travel the 100 yards of the right subtrack and to strike Victim. These needs recur in the actual scenario (the actual passage of the trolley down the 100-yard left subtrack is a counterpart of it’s fallback journey), so no fallback needs are cancelled. Neither does the flipping meet any needs; it has no counterpart event in the fallback scenario. Again, the flip neither meets nor cancels any fallback needs, so E’s need for C is artificial. Is there any more natural fact than the fact that the fallback track is untraveled that would make this need artificial? Again, none that come to mind.²⁷ (DF) says therefore, that the flip is not a cause of Victim’s crushing, which is intuitively the right result.

²⁷ In all subsequent cases, I shall omit discussion of whether a more natural fact exists that would change the verdict of (DF). I have satisfied myself that in all examples detailed here, no such fact readily presents itself. Interested readers may check that this is indeed the case.



$$\begin{aligned}
 F &= 1, L=F \\
 R &= \sim F \\
 V &= L \vee R
 \end{aligned}$$

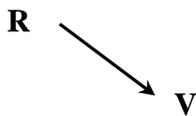
Counterparts : $\langle L, R \rangle$

$$\begin{aligned}
 \text{FAN} &= \{R\} \\
 \text{GAN} \mid R=0 &= \{F, L\}
 \end{aligned}$$

$$\begin{aligned}
 (\text{DF}) &= \underline{\text{No}} \\
 \text{Intuition} &= \underline{\text{No}}
 \end{aligned}$$

Is F a cause of V?

Fallback



Actual

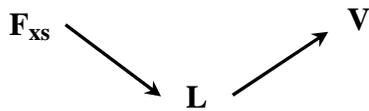


Figure 3.4. Flip: Annotated Causal Graph

3.7 OBJECTIONS TO THE SUFFICIENCY OF (DF)

There appear to be cases of *de facto* dependence between events that one would not normally consider to be related as cause and effect; while the candidate cause event meets or cancels a fallback need, it would be counterintuitive to call it a cause.

One may generate counterexamples to the sufficiency of dependence by starting with examples that are intuitively not cases of causation (for which simple counterfactual dependence fails but *de facto* dependence holds) and then tweaking the facts of the case so that the theoretical verdict changes, while our intuitive judgment remains unchanged. One heuristically fertile

method is to start with Flip (for which everyone agrees that the flipping is not a cause of Victim's crushing) and successively introduce a series of variants of this case in which our intuition is still firm that the flipping is not a cause, but where the flipping either meets or cancels some fallback need.

Let us first take an example in which some fallback need is cancelled. One can generate such examples by simply adding an extra need to the fallback path. As we have already seen, if we make the fallback path 50 yards longer, the flipping cancels the need for the trolley to travel the extra 50 yards.

3.7.1 First Counterexample to the Sufficiency of (DF): Flip 1cm

Let us consider a second variant of Flip:

Flip 1cm: The untraveled fallback track is only one centimeter longer than the actual traveled track.

In this example, one is not inclined to judge the flipping to be a cause of the Victim's crushing. The verdict of the (DF), however, is that the flipping *is* a cause, in virtue of it canceling the need for the trolley to successfully negotiate the extra centimeter of track. Next consider a variant in which the fallback track is several miles long, is in very poor condition and winds through hazardous country. Let the actual subtrack be of *zero* length: i.e. the victim dies *immediately* after the flip. In this case, one's initial intuition is that the flipping is a cause of the death. One suspects that this intuition arises, at least in part, because of the vagaries associated with the longer fallback path. It seems that we know for sure that the victim would get crushed if the switch were flipped, whereas in the fallback scenario, his death seems much less certain: who knows what risks (breakdowns, derailments, etc.) the trolley would encounter on its long and tortuous path? It is plausible that it is these *risks*, rather than the extra length per se, that is the source of the intuitive judgment that the flipping counts as a cause. In the case where the fallback track is only one centimeter longer, there is no reason to suppose that there should be any extra

risks associated with the slightly longer path. One would consider the possibility of the trolley experiencing some mishap on the last centimeter of its journey to be entirely negligible.

Yablo, it seems, agrees with these intuitions regarding the canceling of minor risks in Flip variants:

Although technically, one fallback need is as good as another, in practice not all such needs are taken equally seriously.

[Auto-Reconnection:] What if [the untraveled] subtrack is normally disconnected? The drawbridge it runs over is kept open to let boats through, except when sensors pick up an approaching [trolley], at which times it routinely and automatically closes. The need that Suzy's action cancels is, let us say, the need for a generally reliable mechanism to work the way it is supposed to. Such a need may be considered too slight to protect the [flip] from charges of artificiality.

As an additional illustration, he offers the following case (originally due to McDermott).

[Catch/Wall:] Suppose I reach out and catch a passing cricket ball. The next thing along in the ball's direction of motion was a solid brick wall. Beyond that was a window. Did my action...cause the ball *not* to hit the window?...Usually we think, "The wall was all set to stop that ball; it could and would have done so easily; the catch gets no credit whatever for relieving the effect of so piddling a need." (2002, p.146-7)²⁸

Let us interpret the above remarks in the following manner: Only if the need that is cancelled would have had a significant chance of not being met in the fallback scenario, do we intuitively judge the flip to be a cause. We do not think that there would have been any significant chance that the need for the automatic reconnection of the fallback track would not have been met. Nor do we think it intuitively likely that the wall would not have stopped the ball. We only judge the flipping to be a cause if the need that gets cancelled is an event for which there is a serious risk that it might not occur in the fallback scenario. In other words, if we think that cancelled need *f*

²⁸ Yablo also suggests that sometimes non-objective factors may also play a role in influencing our judgments (and may do so in this particular case) about when a cancelled need is significant.

might not have occurred in the fallback scenario, thereby making the fallback path to be significantly *unreliable*, then we judge the flipping to be a cause. When the fallback path is judged to be pretty reliable, we do *not* judge the flipping to be a cause.

Suggested Repair to (DF): (U)

Suppose that in light of the above considerations we were to decide to modify (DF) by offering the following repair:

(U) Cancelled fallback needs must have had a significant chance of not being met in the fallback scenario (making the fallback path significantly unreliable²⁹).³⁰

Let the conjunction of (DF) and (U) be designated (DFU), the “U” standing for “Unreliable”. (DFU) appears to be a significant improvement on (DF), as now the theory delivers the same verdict as our intuitions for cases in which we do not judge there to be any extra risk associated with the fallback path.

Counterexample to (DFU): Trainee and Meteorite

Unfortunately, (DFU) fails miserably for many cases of preemption, and the unreliability condition (U) cannot therefore be a criterion that we employ in making our intuitive judgments in these cases.

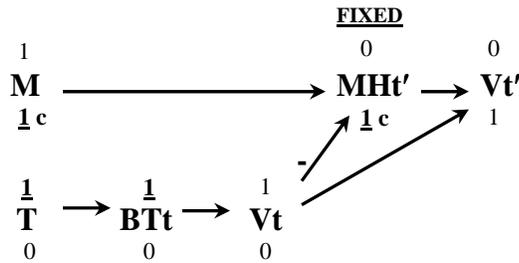
Counterexamples can be generated at will, by simply constructing suitable preemption cases in which canceling takes place, but the fallback path is reliable. Consider the following case of late preemption:

Trainee and Meteorite: A large meteorite is about to hit Victim. Trainee takes out his vaporizing gun and shoots, vaporizing Victim a split-second before the meteorite would have hit him, had he not been vaporized by Trainee. If Trainee had not fired, the meteorite would have vaporized Victim.

²⁹ Of course, “significantly” is somewhat vague, but this may not be a disadvantage. In fact, if it could be shown that our intuitions are similarly vague, it is desirable that the theory match this vagueness in our intuitions.

³⁰ In this and all subsequent suggested repairs to (DF), all of the modifications will be expressed as alternatives to condition (2*) of (DF*).

Let T=Trainee Shoots; BTt=Vaporizing Bullet Hits Victim at time t; M=Meteorite en-route to Victim; MHt' =Meteorite Hits Victim At Time t'; Vt=Victim Vaporized at t; Vt' = Victim Vaporized at t'.



$$\begin{aligned}
 T &= 1, BTt=T \\
 M &= 1, MHt' = M \& (\sim Vt) \\
 Vt &= BTt, Vt' = MHt' \vee Vt
 \end{aligned}$$

Counterparts : None

$$\begin{aligned}
 \text{FAN} &= \{M_c, (MHt')_c\} \\
 \text{GAN} |MHt'=0 &= \{T, BTt, Vt\}
 \end{aligned}$$

Is T a cause of V?

$$\begin{aligned}
 \text{(DF)} &= \underline{\text{Yes}} \\
 \text{(DFU)} &= \underline{\text{No}} \\
 \text{Intuition} &= \underline{\text{Yes}}
 \end{aligned}$$

Fallback



Actual

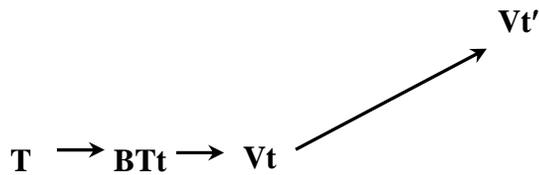


Figure 3.5. Trainee and Meteorite: Annotated Causal Graph

Unquestionably, as with Trainee and Supervisor, our intuition is that Trainee's shooting is the cause of Victim's vaporization. While the vaporization does not depend on Trainee's shooting, it does depend on it if we hold fixed the fact that the meteorite does not strike Victim while he is alive. Although Trainee's shooting cancels the need for the meteorite to hit Victim, we judge it to

be overwhelmingly likely that the meteorite *would* have hit Victim had Trainee not shot; hence the fallback path to Victim’s vaporization is extremely reliable. According to (DFU), however, the cancellation of this need is insufficient to make Trainee’s shooting a cause. Note that in cases of preemption such as this, we do *not* make the intuitive judgments that Yablo claims we make in McDermott’s cricket ball case. We would *never* say: The meteorite was all set to vaporize Victim; it could and would have done so easily; Trainee’s shooting gets no (causal) credit whatever for canceling such a piddling need.

This kind of counterexample to (DFU) arises because our intuitions about preemption cases such as these are simply *not sensitive* to the reliability of preempted backup cause. Indeed, it is plausible to say that our intuitive judgments in preemption cases pay scant (if any) attention to the preempted cause. Instead, it appears that we focus our attention only on the (intrinsic) relations between the preemptor and the effect.

Thus, while it may accord with our intuitions to rule that Flip variants in which the fallback route is reliable are *not* cases of causation, we judge preemptors to be causes *whether or not* their respective fallback routes are taken to be reliable. Hence the addition of condition (U) to (DF) fails.

Table 2 below summarizes our results so far.

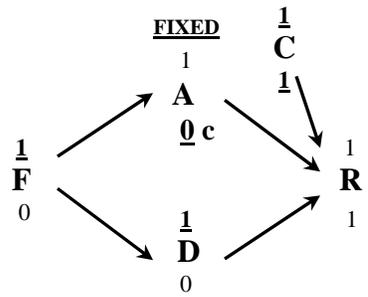
Table 2. Theoretical Verdicts of (DF)

| Case | Intuition | (DFU) | Fact G to be held fixed | FB need met by C? | FB needs cancelled? | FB path unreliable? |
|----------------------|-----------|-------|------------------------------|-------------------|---------------------|---------------------|
| Bomb | N | N | Bomb explodes | N | N | N |
| Flip | N | N | Right subtrack untraveled | N | N | N |
| Flip 1cm | N | N | " | N | Y | N |
| Auto-Reconnect | N | N | " | N | Y | N |
| Catch/Wall | N | N | Ball doesn’t hit wall | N | Y | N |
| Trainee & Supervisor | Y | Y | Supervisor doesn’t shoot | Y | N | N |
| Trainee & Meteorite | Y | N* | Meteorite doesn’t hit Victim | N | Y | N |

* = Intuitively *incorrect* verdict, i.e. a counterexample to (DFU)

3.7.2 Second Counterexample to the Sufficiency of (DF): Two Assassins

Suppose that one is willing to swallow (DF)'s counterintuitive verdicts that flipping is a cause of Victim's crushing in Auto-Reconnect and Flip 1cm. A further problem then arises for (DF) in the form of a surprising objection to our earlier theoretical verdict for Two Assassins. It could be argued that the fallback needs in this case *also* include Assistant's *not* shooting (because in the fallback scenario, even though Captain gives no order, the glowing report still depends on Assistant's not shooting: if Assistant *had* shot in the fallback scenario, then Victim – having not ducked – would have been killed). This dependence on Assistant's not shooting would appear to be *cancelled* by Captain's yelling "Fire!". In the actual scenario, it does not matter to Victim's health whether Assistant fired or not, given that Victim would still have ducked. Notice that the counterfactual must not "backtrack" here: if Assistant hadn't fired, it is *still* the case that Captain would have yelled "Fire!", and so overhearing this order, Victim would still have ducked, whence Assistant's actions are irrelevant. Perversely, the Captain's yell relieves the effect of the fallback need for Assistant not to shoot! And in virtue of this canceling, (DF) rule that Captain's yelling "Fire!" is a cause of Victim's glowing report.



$F = 1, D=F, A=F$
 $C = 1$
 $R = (D \vee (\sim A)) \ \& \ C$

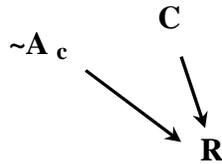
Counterparts : $\langle C, C \rangle$

FAN = $\{C, \sim A_c\}$
GAN $|A=1$ = $\{C, F, D\}$

Is F a cause of R?

(DF) = Yes
Intuition = No

Fallback



Actual

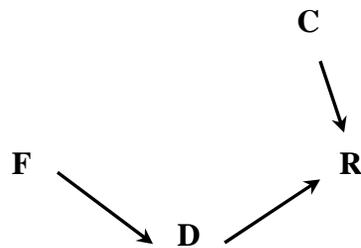


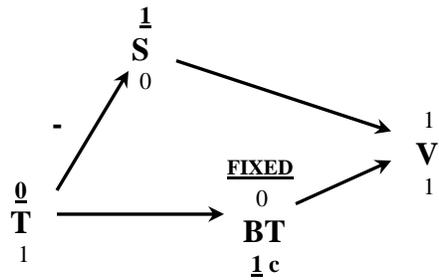
Figure 3.6. Two Assassins Revisited: Annotated Causal Graph

(DF) is therefore unable to deliver the intuitively correct verdict for one of the main types of counterexample that it was expressly constructed to handle! This is something of an embarrassment for the advocate of (DF).

3.7.3 Third Counterexample to the Sufficiency of (DF): Refusal to Shoot

Some variants of early preemptions in which the preempting cause is an *omission* appear to be counterexamples to (DF).

Refusal to Shoot: Supervisor declares that he will kill Victim by activating his action-at-a-distance (AAAD) vaporizing weapon unless Trainee immediately vaporizes Victim with a vaporizing bullet (BT). Trainee refuses. Supervisor activates his weapon and Victim is vaporized (V).



$$\begin{aligned} T &= 1, BT=T \\ S &= \sim T \\ V &= BT \vee S \end{aligned}$$

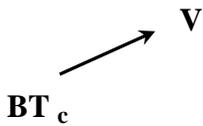
Counterparts : None

$$\begin{aligned} \text{FAN} &= \{BT_c\} \\ \text{GAN} |_{BT=0} &= \{\sim T, S\} \end{aligned}$$

Is $\sim T$ a cause of V?

(DF) = Yes
Intuition = No

Fallback



Actual

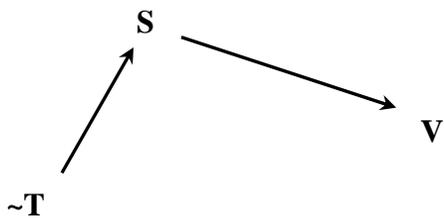


Figure 3.7. Refusal to Shoot: Annotated Causal Graph

Was Trainee’s refusal a cause of Victim’s vaporization? Intuitively, it would seem not. However, if we hold fixed the fact that Trainee’s vaporizing bullet did not hit Victim, then Victim’s vaporization *does* depend on Trainee’s refusal. This is because Supervisor would not have activated his weapon had Trainee not refused. And given that Trainee’s bullet never hits Victim, there is no danger to Victim from either Trainee or Supervisor. In the fallback scenario, the vaporization would have depended on Trainee’s bullet hitting Victim. This need is cancelled in the actual scenario (because Supervisor’s gun has no need of old-fashioned “local” bullets) and hence (DF) delivers the verdict that Trainee’s refusal is a cause of Victim’s vaporization. (DFU), incidentally, delivers the intuitively correct verdict: since the fallback route was (presumably) reliable, the canceling of the need for Trainee’s bullet is not enough for his refusal to count as a cause. To preempt any thoughts of resurrecting (DFU), however, let us replace Trainee’s shooting by Trainee’s stabbing Victim with a fork – a rather unreliable method of killing, but one which, nevertheless, on the occasion in question happened to kill Victim. In this variant, Trainee’s not stabbing would (counterintuitively) count as a cause, according to (DFU).

3.8 OBJECTIONS TO THE NECESSITY OF (DF)

One can construct counterexamples to the necessity of (DF) by starting from clear cases of causation for which (DF) gives the right verdict and adjusting the facts of the case so that the theoretical verdict of (DF) changes, but our intuitions remain the same.³¹ Preemptions are useful starting points.

3.8.1 First Counterexample to the Necessity of (DF): Double Backup³²

Begin with the Trainee and Supervisor example and add in a *second* backup supervisor. If Trainee had not fired, *and* Supervisor 1 (S1) had also not fired (for whatever reason), then Supervisor 2 (S2) *would* have, and Victim would have bled to death (V). But if Trainee had not

³¹ We have already seen that preemptions for which the fallback path is judged to be reliable (either by default or by stipulation) will constitute counterexamples to the necessity of (DFU).

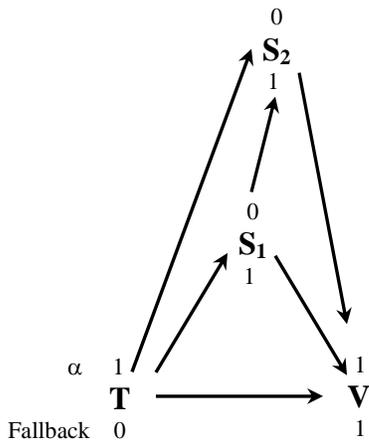
³² Thanks to Cian Dorr for bringing this example to my attention.

fired and Supervisor 1 had fired, then Supervisor 2 would not have. In the fallback scenario, because of the redundancy due to S2, V does *not* now depend on S1's shooting. Hence there are *no* fallback needs. Trainee's shooting therefore *cannot* count as a cause of Victim's bleeding to death since there are no fallback needs for it to meet! Yablo acknowledges these difficulties and admits that (DF) only works as an account of causation where redundancy is one level deep.³³

However, a simple repair suggests itself. What we need is a strategy that will allow us to remove the dependency-breaking effect of *each* redundant backup. By performing a *series* of holding fixed maneuvers, we can remove the backups one at a time. One way to achieve this is by redefining what a fallback need is. Rather than defining a fallback need to be an event on which the effect depends on *simpliciter* in the fallback scenario, let them be the events the effect *de facto* depends on in the fallback scenario. Call this revised version of (DF), in which fallback needs are defined as the events the effect *de facto* depends on in the fallback scenario, '(DF**)' .³⁴ In the fallback scenario for Double Backup, the effect *does de facto* depend on S1's shooting, holding fixed the (non-actual) 'fact' that S2 doesn't shoot (the fallback scenario represents an early preemption of S2 by S1). Let us examine in more detail why this is so. (It may be helpful for the reader to repeatedly refer to figure 3.8 during the subsequent discussion). Note that the fallback scenario can be thought of having its *own* 'second-level' fallback scenario. This is labeled 'Fallback 2' in figure 3.8. In the second-level fallback scenario, the second-level fallback needs are (according to our new definition of fallback needs) the events the effect *de facto* depends on. In the second-level fallback scenario, the effect *de facto* depends on S2's shooting, holding fixed the null fact ϕ (the degenerate case). Hence S2's shooting is a second-level fallback need, which is *met* by S1. Hence V does *de facto* depend on S1 in the fallback scenario, and therefore *does* count as a fallback need after all. Trainee's shooting will therefore be judged, correctly, by (DF**) to be a cause of V, since V depends on Trainee's shooting holding fixed the fact that neither S1, S2, nor S3 shoot, and it also meets the fallback need for S1's shooting.

³³ Personal communication, 2005.

³⁴ Thanks to Brian Epstein for useful discussion of this strategy and for prompting me to think about whether an infinite regress threatens. (It doesn't).



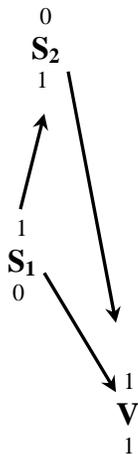
Is T a cause of V?

$T = 1$
 $S1 = \sim T$
 $S2 = \sim T \ \& \ (\sim S1)$
 $V = T \vee S1 \vee S2$
Counterparts : $\langle T, S1, S2 \rangle$

FAN = ϕ
GAN | $S1=0$ and $S2=0$ = $\{T\}$

(DF) = No
Intuition = Yes

Fallback



$S1 = 1$
 $S2 = \sim S1$
 $V = S1 \vee S2$
Counterparts : $\langle S1, S2 \rangle$

FAN₂ = $\{S2\}$

(DF)** = Yes
Intuition = Yes

Second-Level Fallback

Actual

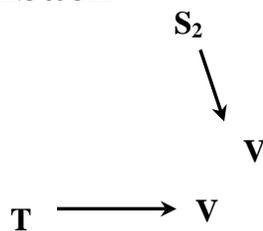


Figure 3.8. Double Backup: Annotated Causal Graph

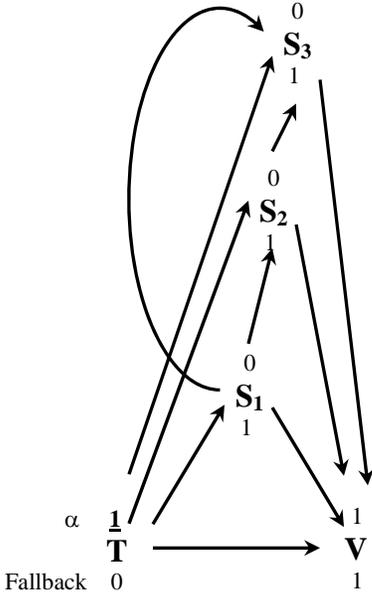
We may add any number of backup Supervisors we like, and (DF**) will still deliver the intuitively correct verdict. Again, the strategy is to remove the dependency-breaking effects of the backups one at a time. (In the section that follows, because the chain of reasoning is rather convoluted, I will number each step). Suppose we add a third Supervisor. Then

1. Trainee's shooting will be judged a cause of V by (DF**) if it meets or cancels a fallback need.
2. A fallback need is an event upon which V *de facto* depends in the fallback scenario (where S1 shoots but S2 and S3 don't).
3. The events upon which V *de facto* depends in the fallback scenario are those that meet or cancel a *second-level* fallback need.
4. A second-level fallback need is an event that V *de facto* depends on in the second-level fallback scenario (where S2 shoots but S3 doesn't).
5. The events upon which V *de facto* depends in the second-level fallback scenario are those that meet or cancel a *third-level* fallback need.
6. A third-level fallback need is an event that V *de facto* depends on in the *third-level* fallback scenario (where S3 shoots).
7. S3's shooting is an event that V *de facto* depends shooting in third-level fallback scenario (holding fixed the null fact ϕ).

Now, working backwards:

8. S3 is a third-level fallback need. (from 7 and 6)
9. The events upon which V *de facto* depends in the second-level fallback scenario are those that meet or cancel S3. (8,5)
10. S2 meets the need for S3 in the second-level fallback scenario.
11. V *de facto* depends on S2 in the second-level fallback scenario. (10,9)
12. S2 is a second-level fallback need. (11,4)
13. The events upon which V *de facto* depends in the fallback scenario are those that meet or cancel S2. (12,3).
14. S1 meets the need for S2 in the fallback scenario
15. V *de facto* depends on S1 in the fallback scenario. (14,13)
16. S1 is a fallback need. (15,2)

- 17. Trainee's shooting meets the need for S1 in the actual scenario.
- 18. Trainee's shooting is a cause of Victim's bleeding to death. (17,1)



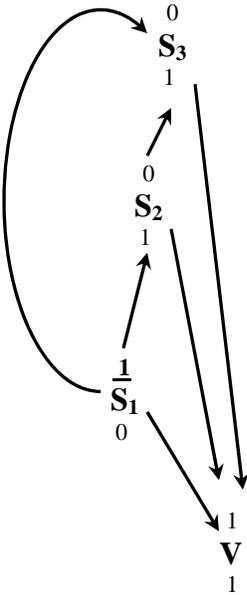
$T = 1$
 $S1 = \sim T$
 $S2 = \sim T \ \& \ (\sim S1)$
 $S3 = \sim T \ \& \ (\sim S1) \ \& \ (\sim S2)$
 $V = T \vee S1 \vee S2 \vee S3$
Counterparts : $\langle T, S1, S2, S3 \rangle$

FAN = ϕ
GAN $|S1, S2, S3=0$ = $\{T\}$

Is T a cause of V?

(DF) = No
Intuition = Yes

Fallback

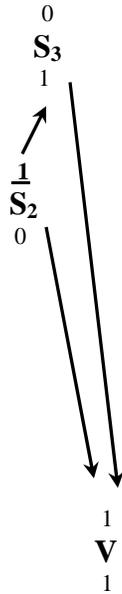


$S1 = 1$
 $S2 = (\sim S1)$
 $S3 = (\sim S1) \ \& \ (\sim S2)$
 $V = S1 \vee S2 \vee S3$
Counterparts : $\langle S1, S2, S3 \rangle$

FAN₂ = ϕ
GAN₂ $|S2=0$ and $S3=0$ = $\{S1\}$

Figure 3.9a. Triple Backup: Annotated Causal Graph

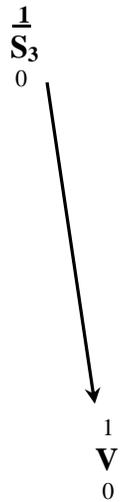
Second-Level Fallback



$S_2 = 1$
 $S_3 = \sim S_2$
 $V = S_2 \vee S_3$
Counterparts : $\langle S_2, S_3 \rangle$

$FAN_3 = S_3$
 $GAN_3 | S_3=0 = \{S_2\}$

Third-Level Fallback



Actual

T \longrightarrow **V**

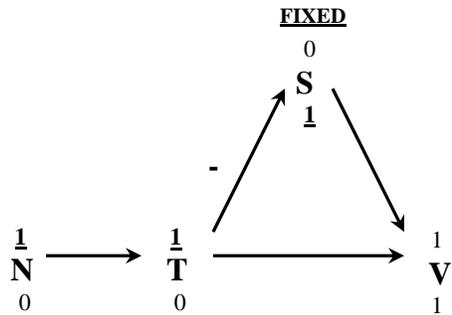
Figure 3.9b. Triple Backup: Second and Third-Level Fallback Scenarios

By redefining fallback need, (DF) is thus able to resist counterexamples in which there is multiple redundancy. This strategy for handling multiple redundancy is not very elegant but it achieves the correct result by brute force. There are, however, a series of further counterexamples to the necessity of (DF), for which (DF**) is no help.

3.8.2 Second Counterexample to the Necessity of (DF): Nudge

It is a fairly straightforward matter to make a *bona fide* cause appear to be piled artificially on top of the fallback needs. Counterexamples can be generated at will by starting from a clear-cut case of causation (e.g. early preemption), and simply “front-loading” further events *before* the preempting cause, on which the preempting cause depends. For example, starting with Trainee and Supervisor, front-load by inserting an earlier event on which Trainee’s shooting depends:

Nudge: Trainee has his Victim lined up in his cross-hairs but has decided not to shoot. Colonel, who is passing by, accidentally nudges Trainee in the back, making him shoot. If he had not nudged him, Trainee would not have fired, and so Supervisor would have.



Is N a cause of V?

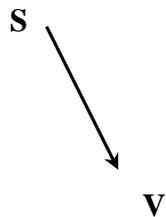
$$\begin{aligned}
 N &= 1, T=N \\
 S &= \sim T \\
 V &= T \vee S
 \end{aligned}$$

Counterparts : $\langle T, S \rangle$

$$\begin{aligned}
 \text{FAN} &= \{S\} \\
 \text{GAN} \mid S=0 &= \{T, N\}
 \end{aligned}$$

(DF) = No
Intuition = Yes

Fallback



Actual

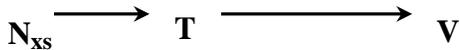


Figure 3.10. Nudge: Annotated Causal Graph

My intuition is that Colonel’s nudge was a cause of Victim’s death. I take it that the nudge would have caused the death in the absence of Supervisor and do not feel that my intuition changes in its presence. Yet holding fixed that Supervisor did not shoot, Victim’s death depends on the nudge. In this case, the putative cause (the nudge), does not meet any fallback need (unlike Trainee’s shooting in Trainee and Supervisor); there is no fallback need of which the nudge is a counterpart. Neither did the nudge cancel the fallback need for Supervisor’s shot; its counterpart

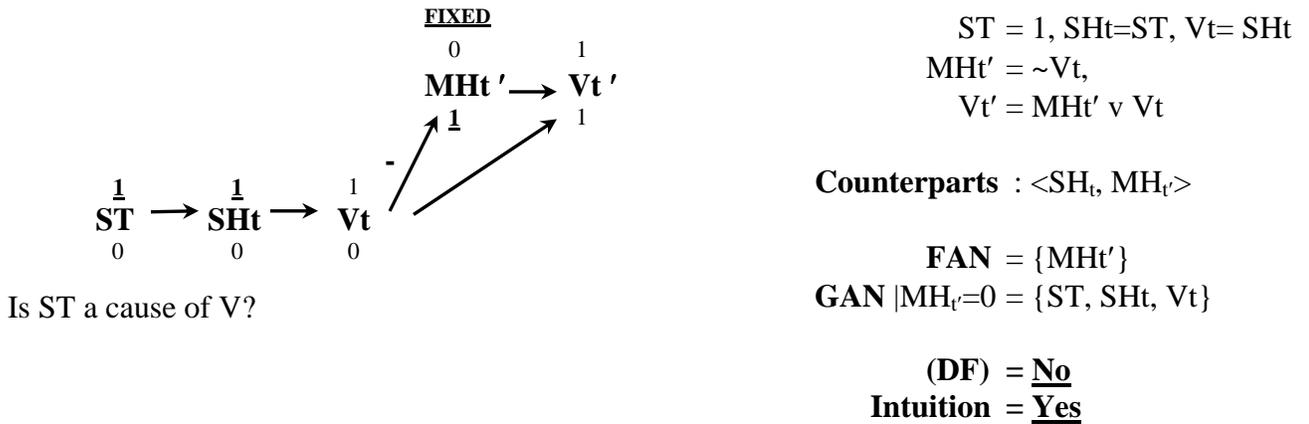
(Trainee's shot) is still required in the actual scenario. Hence (DF) rules that the nudge is not a cause of Victim's death. Yet intuitively, it seems that the nudge *should* count as a cause.³⁵

3.8.3 Third Counterexample to the Necessity of (DF): Small Meteorite and Suzy

A second related method of generating counterexamples to the necessity of (DF) is, rather than inserting earlier events, to modify cases of preemption by removing the putative cause's counterpart. The putative cause is then no longer able to meet the fallback need for that counterpart. For example:

Small Meteorite and Suzy: This example is the same as Billy and Suzy, except that we take away Billy's throw and replace it with a small incoming meteorite that exactly replicates the trajectory of Billy's rock.

³⁵ It is not even necessary to make Trainee's shooting dependent on the nudge. If we state that Trainee would still have fired, even if he hadn't been nudged, only a little later, it is my intuition that the nudge still counts as a cause. In this case, it would hasten Victim's bleeding to death. For me, considerations of momentum transfer trump the lack of dependence.



Fallback

$$MHt' \rightarrow Vt'$$

Actual

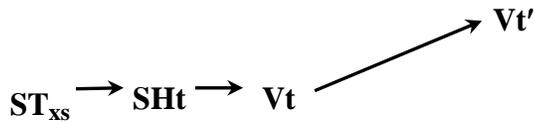


Figure 3.11. Small Meteorite and Suzy: Annotated Causal Graph

Suzy's throw now has no counterpart in the fallback scenario, and hence it meets no fallback need. Nor does it cancel any, for the fallback need met by the meteorite hitting the bottle is met by its counterpart in the actual scenario: Suzy's rock hitting the bottle. Hence Suzy's throw neither meets nor cancels any fallback needs; it is artificially piled on top of the fallback needs. It is not therefore a cause of the bottle's shattering, according to (DF). But the intuition that Suzy's throw is a cause remains as solid as it always is in cases of late preemption. Note that this case differs from slightly from Trainee and Meteorite, where the need for the meteorite was cancelled by Trainee's shooting, and in virtue of which Trainee's shooting was judged by (DF) to be a cause of Victim's vaporization. In Small Meteorite and Suzy, because the small meteorite is a counterpart of Suzy's rock (in size, mass, shape, momentum, etc.), the need for it is *not*

cancelled. And because Suzy's throw does not obviously meet any fallback need, (DF) rules that Suzy's throw is *not* a cause, which is counterintuitive; this constitutes a blow to (DF).

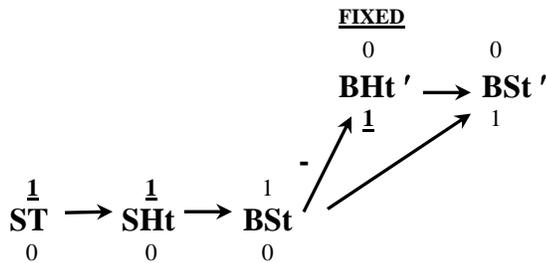
It might be objected that there is an event somewhere in the meteorite's past that originally gave the meteorite its earth-directed momentum. Suzy's throw is either a counterpart of this event, or if it is sufficiently different, Suzy's throw *Cancels* this event.

I have two responses to this objection. The first is to make use of Yablo's *own* rule that neither the actual nor the fallback needs can include events *before* the branch point. This rule is included, one presumes, in order to rule out events in the past that are not relevant to the causal status of the preemptor and the preempted cause. Unfortunately, it excludes many events that are held by the de facto-ist to be highly relevant to C's causal status. With regard to Small Meteorite and Suzy, the rule does not then permit the putative event which gave the meteorite its momentum to be included in the effect's fallback needs, if that event occurred before Suzy's throw. This "post-branch point" rule is highly problematic for (DF). Even in the standard case of late preemption (Billy and Suzy) if Billy throws *before* Suzy, (but, for example, is slightly further away, or throws less hard than Suzy does, so that Suzy's rock still hits the bottle first), Billy's throw will not be among the shattering's fallback needs! (DF) will deliver this verdict in any example in which the preempted cause predates the preempting cause.

My second response to the objection to Small Meteorite and Suzy is to use the heuristic trick of switching reference frames, so that the momentum of the meteorite is eliminated. Then we no longer have to worry about the putative momentum-producing event in the meteorite's history. In the reference frame of the meteorite, the meteorite is stationary, with the bottle hurtling towards it. If Suzy had not thrown, her rock would never have hit the bottle. The shattering's actual needs are Suzy's throw, the motion of the bottle and Suzy's rock hitting the bottle. The fallback needs are the motion of the bottle and the meteorite's hitting the bottle (of which Suzy's hitting is a counterpart). Thus Suzy's throw neither meets nor cancels any fallback needs. An even clearer way of making the same point is via the following (somewhat) analogous case:

Swinging Bottle: A bottle is suspended from a tree branch by a piece of string and is swung towards Billy's stationary rock, which is also suspended from the same branch. However, just before the bottle collides with Billy's rock, Suzy's rock strikes the bottle,

and the bottle shatters. Had Suzy not thrown, the bottle would have shattered on colliding with Billy's rock.



$$ST = 1, SHt = ST, BSt = SHt$$

$$BHt' = \sim BSt$$

$$BSt' = BHt' \vee BSt$$

Counterparts : $\langle SHt, BHt' \rangle$

FAN = {BHt'}

GAN |BHt'=0 = {ST, SHt', BSt}

Is ST a cause of V?

(DF) = No

Intuition = Yes

Fallback

$$BHt' \rightarrow BSt'$$

Actual

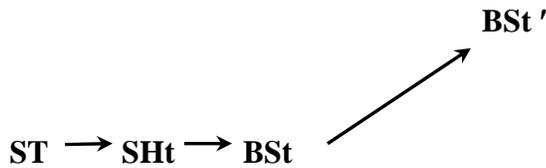


Figure 3.12. Swinging Bottle: Annotated Causal Graph

As in the previous case, Suzy's throw neither meets nor cancels any fallback needs, and hence (DF) says, counterintuitively, that it is not a cause of the bottle's shattering.

3.8.4 Fourth Counterexample to the Necessity of (DF): Billy and Suzy Deflect

This reference frame switching trick can also be used in cases of late cutting:

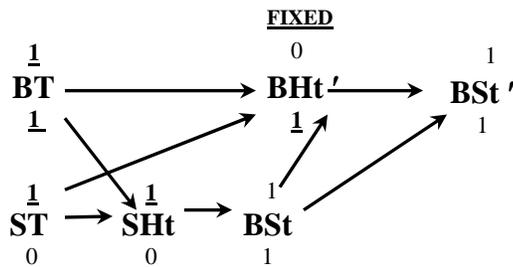
Billy and Suzy Swinging Bottle Deflect: Suzy's throw is a little off-target. But it happens to collide with Billy's stationary suspended rock, and is deflected into the path of the swinging bottle, which shatters.³⁶

The fallback needs are for the swing and for Billy's rock to hit the bottle. The actual needs are for Suzy's throwing, the swinging of the bottle, Suzy's rock's collision with Billy's rock and Suzy's rock's hitting the bottle. All of the fallback needs recur in the actual scenario, and Suzy's throwing meets *none* of them. (DF) therefore rules that Suzy's throwing is not a cause of the bottle's shattering, which is again highly counterintuitive.

In fact, we do not even need to make the reference frame switch. Consider:

Billy and Suzy Deflect: Assume that Billy and Suzy both throw. Billy's throw is on target but Suzy's is not. Again, Suzy's rock takes a deflection from Billy's rock onto the exact trajectory that Billy's rock would have followed, and goes on to shatter the bottle. Billy's rock is knocked off its original on-target trajectory and continues along the trajectory that Suzy's rock would have taken, had the collision not occurred.

³⁶ Adapted from Yablo (2004).



$ST = 1, SH = ST \ \& \ BT$
 $BT = 1, BH = BT \ \& \ (\sim ST) \ \& \ (\sim BSt)$
 $BSt = SHt$
 $BSt' = BHt' \ \vee \ BSt$

Counterparts : $\langle BT, BT \rangle, \langle SHt, BHt' \rangle$

FAN = {BT, BHt'}

GAN |BH=0 = {BT, SHt, ST, BSt}

Is ST a cause of BS?

(DF) = No

Intuition = Yes

Fallback



Actual

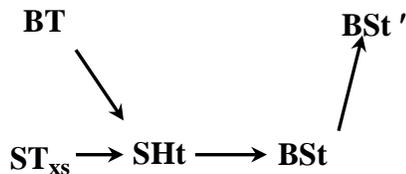


Figure 3.13. Billy and Suzy Deflect: Annotated Causal Graph

The fallback needs are for the swing and for Billy’s rock’s hitting the bottle. The actual needs are for Suzy’s throwing, Suzy’s rock’s colliding with Billy’s rock, Suzy’s rock’s hitting the bottle, and also for Billy’s throwing (since if he had not thrown, there would have been no mid-course correction to Suzy’s off-target throw). Billy’s rock’s hitting the bottle and Suzy’s rock’s hitting the bottle are counterparts. Suzy’s throw, however, should not be judged the counterpart of Billy’s throw. The counterpart of Billy’s fallback throw is Billy’s actual throw! Billy’s throw and Suzy’s throw are not counterparts in terms of momentum: Billy’s throw is an on-target throw, while Suzy’s is an off-target throw. If it is objected that Suzy’s throw meets the need for putting

Billy's rock in motion, then it can be replied that this need recurs in the actual scenario: Billy's rock still needs to be put into motion so that it can collide with Suzy's. The need to put Suzy's rock into motion is an *additional* need, over and above this. So Suzy's throw neither meets nor cancels any fallback needs, and by (DF), is not a cause. Even without working through the technical details of this case, we can see instinctively that it is likely to be a counterexample to (DF). Using the original insight that events which are piled on artificially over and above fallback needs should not count as causes, we can quickly see that Suzy's throw is piled on artificially in this manner: everything that had to happen in the fallback scenario still has to happen in the actual scenario (in order to deflect Suzy's rock). Suzy's throw is over and above these needs.

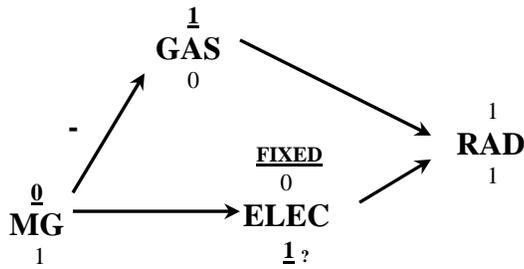
3.9 (DF) IS UNDECIDABLE

3.9.1 The Counterpart Problem

There appear to be cases for which it is not entirely clear what verdict (DF) actually delivers. It is usually considered important that a theory deliver determinate verdict, i.e. that it should be decidable. However, this undecidability may be considered a theoretical virtue, if we intuitively find the same cases indeterminate. If, in addition, the theory appears to be a plausible account of the steps that were actually followed (probably not explicitly) in reaching the intuitive verdict, this would be an added bonus, leading us to judge the theory even more favorably. While one might not necessarily require a theory to explain our intuitions (rather than just rationally reconstruct them, or get the extension right), it seems like a good thing if the theory is able to accomplish this too. Let us look at some cases in which it is not obvious what verdict (DF) gives.

Power Failure: A town's main generators fail, and in consequence, the electricity supply to the library is interrupted and the central heating radiators shut down. The library has a back-up system in place that is activated if a sensor detects that the main generator's dynamo is not rotating. If the generators fail, the central heating system switches over to

a back-up gas supply, which heats the water for the radiators (in place of the electric water heaters). A few minutes after the failure of the main generators, the radiators in the library are back on. (Adapted from Hitchcock, 2003).



MG = 0, ELEC=MG
 GAS = ~MG
 RAD = ELEC v GAS

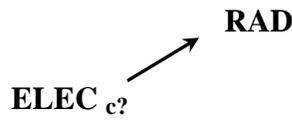
Counterparts : ?

FAN = {ELEC?}
 GAN |E=0 = {~MG, GAS}

Is ~MG a cause of RAD?

(DF) = ?
 Intuition = No

Fallback



Actual

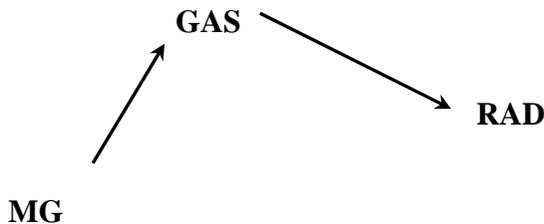


Figure 3.14. Power Failure: Annotated Causal Graph

Does the failure of the main generators cause the radiators to be on? One is tempted to say no (as we did for Refusal to Shoot, perhaps because of the fact that in general, power failures are not conducive to the radiators working properly). The verdict of (DF), however, depends on whether we view the gas supply as a counterpart of the electricity supply. In the fallback scenario the radiators' being on depends only on the electricity supply getting through to the library. If the

electricity supply had been interrupted, the gas backup would not have kicked in, because the main generators would still be working. We must not backtrack this counterfactual. Note that in the fallback scenario, the radiators' being on does not depend on the generators' working normally; if they failed the backup gas would keep the radiators on (we must allow this counterfactual to foretrack, as normal).

In the actual scenario, where the main generators *do* fail, given that we hold fixed the fact that the electricity supply failed, the radiators' being on depends on the failure of the main generators (in order to activate the gas backup), and, of course, on the gas supply.

If the gas supply is a counterpart of the electricity supply then the dependence on the failure of the generators is artificial, as the failure neither meets nor cancels the fallback need (electricity supply to the library). If they are not counterparts, then the fallback need for the electricity is cancelled, and the failure of the generators is a cause of the radiators being on. On the one hand, they appear to be rather different, with heat being produced by different mechanisms. On the other hand, it could be argued that they meet the same need (for example, the need for an energy source to heat the water). This is one instance of what might be termed "the counterpart problem": it is not entirely clear precisely what conditions must be met in order for events to be counterparts.

3.9.2 The Halting Problem.

It is not always easy to find a fact that puts E in need of C. For example, the fact that meteorite does not hit Victim while alive in Trainee and Meteorite is not perhaps the first thing that would come to mind. The problem is that there is nothing to tell us *when* we should give up our search. There is no finite algorithm that will lead to a definitive verdict. Two individuals with different levels of creativity and persistence might attribute different verdicts to (DF). This problem also besets other Holding Fixed theories, (e.g. Hitchcock 2001, Pearl 2001, Woodward 2003, etc.). Different verdicts are generated depending on which variables are included, and arguments must be given for the appropriateness of the chosen model. For example, if the variable "Meteorite Hits Victim at time t" is included in the causal graph for Trainee and Meteorite, the theory will deliver the right verdict. If it is not included, the theory will give the wrong verdict. As with

(DF), there is no systematic, finite procedure that can be followed in order to generate a verdict. It depends on which events or variables one has considered as possible Gs to hold fixed. This is troubling, since our intuitions do not obviously depend on such considerations.

3.10 CONCLUSIONS

(DF) faces serious objections to its sufficiency. The suggested repair to (DF), (DFU) could not deliver the intuitively correct results for all of the variants of Switching and Preemption cases presented. (DF) also faces objections to its necessity. (DF) does not just struggle to cope with the rather unnatural and exotic counterexamples presented in this chapter. It is also unable to handle some of the more familiar, mundane canonical counterexamples such as early preemption, that other counterfactual theories handle easily. While (DF) appears to make progress in excluding some of the hard cases, the extra technical apparatus it introduces lets back in some of the easy cases. The principal reason that (DF) fails is that it is too easy to take an example that is intuitively clear, to slightly alter it by adding or subtracting events from either the actual or fallback routes in such a way that the theoretical verdict changes, while leaving our intuitions unaltered.

My suspicion is that the Holding Fixed strategy may be something of a red herring. It is certainly a very effective method of getting the right result for early and late preemption. But for almost all the other canonical types of example, it creates all manner of complications. If an *alternative* method of treating preemption correctly were available, then we wouldn't need to use the Holding Fixed strategy at all. In the next chapter, I discuss pluralistic treatments of causation that have recently been offered by Hitchcock (2003) and Hall (2004). Hall suggests separating causation into two relations, which offers the possibility of handling preemption without explicit reference to counterfactuals. Given the problems with (H1), (H3) and (DF) that have been discussed in this chapter and in chapter two, Hall's suggestion is an attractive one.

4.0 PLURALISTIC THEORIES OF CAUSATION

4.1 INTRODUCTION AND OVERVIEW

In chapters two and three, I argued that even the most impressive extant counterfactual analyses of token causation have serious problems. Given that other univocal analyses fare no better, one response would be to simply abandon the search for a univocal theory. Recently, Hall (2004), Hitchcock (2003) and Cartwright (1999, 2004, forthcoming) have offered pluralistic accounts of causation.

In section 4.2 I begin with a brief account of some uncontroversial forms of pluralism about causation that do not threaten the univocalist. In section 4.3, I discuss Hitchcock's distinction between 'net' and 'component' causes. No single analysis, it is claimed, can (or should attempt to) accommodate both subconcepts. Instead, each requires a separate analysis. I will argue that while it is useful for certain practical purposes to introduce, and provide stipulative definitions for the notions of net and component cause, *we don't need to do so* for the purposes of analyzing our ordinary concept of token causation.

Having dismissed this form of pluralism, in section 4.4 I discuss Hall's pluralistic theory of causation (Hall 2004). Hall proposes that there are *two* concepts of causation: production and dependence. Hall's pluralistic theory, since it does not need to appeal to the Holding Fixed maneuver that was problematic for (H1), (H3) and (DF), has the great advantage of not ruling in Switching and Self-Canceling Threats as genuine cases of causation. This is an important advance on univocal theories. I present several counterexamples to Hall's theory that exhibit neither production nor dependence, but which we intuitively judge to be causation.

4.2 UNCONTROVERSIAL TYPES OF PLURALISM

4.2.1 Plurality of Causes of an Event

Events frequently have more than one cause. It seems perfectly correct, for example, to list the recent drought and the tossed cigarette as causes of the forest fire. It also seems correct to label the presence of oxygen in the air as a cause, though this is much less prominent than the tossed cigarette. It is common to draw some kind of distinction between proximate/triggering causes and standing or background conditions, and typically the former are more readily identified as causes. It is often natural to highlight proximate causes that correspond to positive events involving a change of state (e.g. the tossing of the cigarette) rather than persisting states (e.g. the presence of oxygen), though this need not be the case. As Hitchcock has noted, pragmatic considerations also play a role in determining which causes are emphasized: the meteorologist naturally focuses on the drought as the cause of the fire, the policeman on the tossed cigarette.³⁷ I take this type of pluralism to be uncontroversial, and one that poses no threat to the univocalist. The counterfactualist, for example, can simply point out that the fire depends on both the drought and on the cigarette tossing, and therefore correctly judge each to be a cause.

4.2.2 Type/Token Plurality

“Smoking causes cancer” is a general or *type-level* causal claim. The relata of type-level causation are kinds of event or properties. Type causation usually concerns some *population* of individuals. By contrast, “Billy’s throwing the baseball caused the window’s shattering” is a singular or *token-level* claim. The relata of token causal relations are usually taken to be events.³⁸ Token causation is also (increasingly) referred to as *actual* causation. I will mention type-level causation only briefly in this dissertation since the bulk of the philosophical literature on causation has been concerned with the conceptual analysis of token causation.

³⁷ Hitchcock 2003, p.5.

³⁸ Arguments have also been given in favor of taking the relata to be facts (Mellor 1995), objects (Dowe 2000), and so on.

Some philosophers have sought to analyze type and token causation in different ways. Elliot Sober (1985), for instance, gives a probabilistic theory of type causation and a physical process theory of token causation. However, type/token pluralism doesn't necessarily entail that separate analyses are required for the type-level locution "C causes E" and the token causal locution "C was a cause of E". The following univocal probabilistic analysis for example, might potentially cover both:

C raises the probability of E, where C and E can be either events, kinds of event or properties.

While probabilistic theories of causation face well-known counterexamples (especially regarding token causal claims), in principle there is no reason why some other univocal theory might not successfully handle both type and token causation.

There are a variety of other distinctions that could be made between different types of causes. For example: proximal vs. distal, direct vs. indirect, and so on. While these distinctions are interesting in their own right, I do not think that they provide any obstacle to the provision of a successful univocal conceptual analysis and I will not discuss them further.³⁹ The type of pluralism I wish to focus on, however, is that which *does* pose a threat to the univocalist.

4.3 HITCHCOCK'S NET AND COMPONENT CAUSES

Hitchcock (2003) proposes a type of pluralism that *is* a threat to the univocalist. He claims that in a variety of cases, individuals disagree with one another in their intuitions about whether or not C is a cause of E. He states that if two individuals have different intuitions, and neither is obviously mistaken, no univocal theory could possibly accommodate both intuitions. Hitchcock writes:

Theories of causation are typically tested by comparing their verdicts with those of intuition. Our survey [of examples] will demonstrate just how inconsistent and imprecise our intuitions are. In the face of these multiple disagreements [over whether "C is a cause

³⁹ Hitchcock (forthcoming) and Godfrey-Smith (forthcoming) provide further reviews of different ways in which one might be a pluralist about causation.

of E” is true for a particular case], it becomes implausible that our intuitive causal judgments are attuned to one single objective relation. No one theory of causation can be expected to fit with all of these intuitions because the intuitions themselves are incompatible (2003, p.9).

In this section, I will discuss the Birth Control Pills case at some length, since this case is pivotal in Hitchcock’s argument for a distinction between ‘net’ and ‘component’ causes. If intuitions are in conflict in this case, I claim that either one of the intuitions is clearly mistaken, or the disagreement is not due to any ambiguity between net and component causation. Whichever is the case, we do not need the concepts of net and component causation. While I am happy say that net and component causation *are* causal relations in some broad sense, I will argue that such a conceptual distinction is not necessary for the purposes of analyzing token locution “C is a cause of E”, or at least that the Birth Control Pills does not by itself force us to do so.

Let us look at the details of the birth control pill example more closely. This case was initially presented by Hesslow (1976) as a counterexample to probabilistic theories of causation. Hitchcock’s presentation of the case is as follows⁴⁰:

Birth Control Pills: Thrombosis, or the forming of blood clots in the arteries, is considered to be one of the most worrisome side-effects of birth control pills. This means, presumably, that the consumption of oral contraceptives *causes* thrombosis. Yet among women who are fertile, sexually active, and otherwise quite capable of becoming pregnant, and who are under 35, non-smokers, and otherwise at low risk of thrombosis, birth control pills *lower* the overall probability of thrombosis. This is because birth control pills are effective preventers of pregnancy, which itself is a significant risk factor for thrombosis.

Hitchcock asks: “Do birth control pills *cause* thrombosis, or do they *prevent* thrombosis?”, and concludes that we feel compelled to say that they both cause *and* prevent thrombosis. He interprets this ambivalence as a symptom of a fundamental ambiguity in our concept of causation. When asked whether some C is a cause of E, we might disambiguate the question in

⁴⁰ Hitchcock (2003, p.11-12).

different ways, he claims. On the one hand, we might understand the question to be asking “Is C a cause of E along some *particular* causal route,” or, on the other hand, “Is C a cause of E overall, when *all* routes are taken into account?” The intuitive answer to the first question may be ‘yes’ and the second ‘no’. Along the route that prevents pregnancy, the component effect on thrombosis is negative. Along the other route, the component effect is positive (presumably mediated by the hormonal content of the pill). Overall, the net effect is negative, at least in the subpopulation described. Since these intuitions are incompatible, no single theory of causation could accommodate them both, so the argument goes. Hitchcock suggests that we distinguish between *net* and *component* causes. This move resolves the apparent inconsistency in our causal judgments, since when two individuals appear to be disagreeing about some particular causal claim, they are not really disagreeing at all. One is making a claim about net causation, the other a claim about component causation.

There are many ways in which net and component causes could be defined. The counterfactualist could say that C is a component cause of E if and only if there is an active route between the two (i.e. if E depends on C while holding all off-route variables fixed at their actual values), and that C is a net cause of E if E depends on C, in an overall sense (i.e. holding nothing fixed). A probability theorist could define a component cause in terms of probability raising along the relevant route, and net cause in terms of overall probability raising. Hitchcock distinguishes between *stage one* pluralism and *stage two* pluralism. Stage one pluralism refers to the plurality of conceptual ‘building blocks’ that one might employ to analyze causation: counterfactuals, probability relations, causal processes, and so on. Stage two pluralism refers to the variety of causal subconcepts that can be defined using these building blocks. Net/component pluralism is of the stage two variety. It is perfectly possible to be a pluralist at stage two without being a pluralist at stage one.⁴¹

I suggest that we do *not* need to distinguish these two senses of causation in cases like Birth Control Pills if our goal is analysis of token causation. If we assume determinism, there should be *no ambivalence whatsoever* over whether or not the taking the pill was a cause of Betty’s thrombosis. Even if we do not assume determinism (i.e. we assume that the case is *irreducibly* indeterministic), it is not obvious that any unclarity of intuition is due to an

⁴¹ Sober (1985), as presented above, is a stage one pluralist.

ambiguity between net and component causation. Hence this example poses no threat to the univocalist.

My arguments in support of these claims are facilitated by representing the counterfactual structure of this case in a causal model (figure 4.1). Let the binary variables **PILL**, **PREG** and **THROMB** represent whether or not Betty takes the pill, becomes pregnant and suffers a thrombosis. To prevent this example from becoming too complex and obscuring the relevant issues, let us make the following two simplifying assumptions: (i) The pill is an effective preventer of pregnancy, and (ii) If Betty had not taken the pill, she would have become pregnant. The relevant structural equations are then: $PILL=1$, $PREG=\sim PILL$. Let us also assume that $P(THROMB|PILL \ \& \ \sim PREG)=0.2$ and $P(THROMB|\sim PILL \ \& \ PREG)=0.3$.

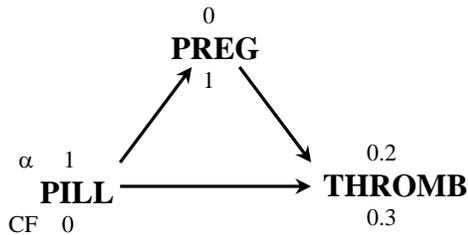


Figure 4.1. Birth Control Pills

It is natural to assume that this example is deterministic, i.e. that no microphysical indeterminacy percolates up to the physiological level at which thrombosis occurs. It is possible that the converse is true, however. My principal concern in this dissertation (as it has been in much of the causation literature) is whether or not token causation can be analyzed univocally in the *deterministic* case and so I will not devote much time to discussing indeterministic cases. Nevertheless, I will make a few comments in section 4.3.2 about the need for the introduction of the concepts of net and component causation in the indeterministic case.

4.3.1 The Deterministic Case

If the Birth Control Pills example is deterministic, there must be hidden variables. Let us assume the structural equation for thrombosis is $THROMB=(PILL\&X) \vee (PREG\&Y) \vee Z$, where X

represents the hidden co-factor for the pill (let us say that X is some genetic predisposition to thrombosis, such that if X occurs, the pill will be sufficient for thrombosis), Y is the hidden co-factor for pregnancy, and Z represents the disjunction of all the other factors on which thrombosis depends in an individual who is neither pregnant, nor takes the pill. The corresponding causal graph is:

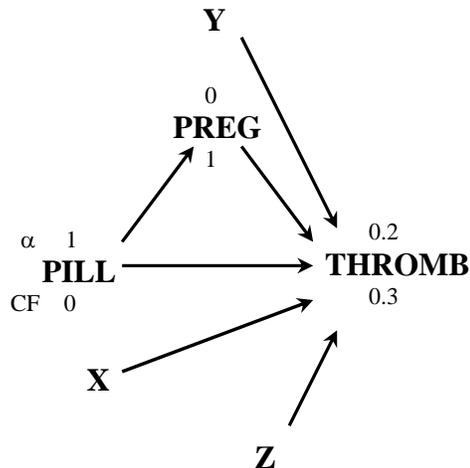


Figure 4.2. Birth Control Pills (Deterministic case)

Now, in this case, I claim there should be *no* debate over whether Betty’s taking the pill was a cause of her thrombosis. In the actual case, both the putative cause (Betty’s taking the pill) and the effect (her thrombosis) are occurrent events. In addition, Betty did not become pregnant. Obviously then, her thrombosis was not caused by her pregnancy. Therefore, either the pill was a cause (in combination with co-factor X) of her thrombosis, or Z was the cause. As long as we are told the relevant facts about whether or not X and Z occurred, there should be no indeterminacy about whether or not Betty’s taking the pill was a cause of her thrombosis. If X occurs and Z doesn’t, the pill was clearly a cause. This is a perfectly straightforward case of preemption; the pill preempts the threat of thrombosis posed by pregnancy. If X doesn’t occur, Z (or rather, one of its disjuncts) was the cause. If we are not told the relevant facts about X and Z, then we should refrain from assigning a truth value to the claim “Betty’s taking the pill was a cause of her thrombosis.” This withholding of judgment does not reflect any genuine semantic indeterminacy, however. The indeterminacy is of an epistemic nature only.

The above discussion is incomplete. It is of course possible that *both* X and Z occur, in which case we would have an instance of symmetric overdetermination. For the canonical examples of symmetric overdetermination in the causation literature, the consensus seems to be that most individuals judge that both putative causes are indeed causes. If Birth Control Pills is analogous to these canonical examples, Betty's pill taking would still be a cause of her thrombosis. Of course, it may be that further information would lead us to reverse this intuition. For example, if it were found that despite taking the pill and having co-factor X, one the physiological processes on which thrombosis normally depends did not occur or was interrupted (and that therefore the pill-thrombosis process did not run to completion) then obviously we would not judge X to be a cause of the thrombosis. But in that case, the causal graph in figure 4.2 and its associated structural equations do not accurately reflect the facts of the case. If it is doubts about whether or not such an interruption occurred that underlie an intuitive reluctance to assign a truth value to the causal claim, again, this indeterminacy is only of an epistemic nature, and should not concern us. In summary, if we are presented with all the relevant facts of the case, an individual should have no trouble in determining whether or not Betty's pill taking caused her thrombosis. Hence there would be no ambivalence to be explained by an appeal to the notions of net and composite causes.

4.3.2 The Indeterministic Case

If the case is indeterministic then things are somewhat different. It is difficult to think about these cases clearly, since we are so accustomed to thinking in terms of deterministic mechanisms. Let us imagine, however, that there is an irreducibly indeterministic 'gap' somewhere in the physiological processes linking Betty's taking the pill to her subsequent thrombosis, which cannot be filled in by any deterministic physiological mechanism. Let us assume, as before, that in the actual world, the pill was effective and that Betty did *not* become pregnant. The relevant details of the indeterministic case then correspond to figure 4.1. Let us in addition stipulate that if Betty were not at risk of becoming pregnant, and were not taking the pill either, her risk of thrombosis would be 0.1. Certainly pregnancy was not a cause of Betty's thrombosis, since by hypothesis, Betty did not become pregnant; as in the deterministic case, any

possible threat from pregnancy has been preempted by Betty's taking the pill. Was Betty's pill taking a cause of her thrombosis? Do we feel ambivalent about this question in the token case, in the way that Hitchcock claims we do in the type case? We clearly don't feel as strongly in the token case that the pill caused Betty's thrombosis as we feel in the type case that the pill is a cause of thrombosis; After all, the probability of thrombosis after Betty had taken the pills was *decreased* to only 0.2, and we know that there was a chance that she would have a thrombosis in the absence of the pill. And we certainly don't feel that the pill *prevented* Betty's thrombosis in the token case, since the thrombosis actually occurred.

My personal view is that in the indeterministic token case, it is simply *mistaken* to view Betty's pill taking to be a cause of her subsequent thrombosis. Consider the sequence of events: Sexually active Betty starts taking the pill, and the threat of thrombosis from pregnancy is removed. At this point, her probability of thrombosis is 0.2. Thereafter, nothing else happens to alter this probability. Then Betty's thrombosis occurs. It is very tempting to view Betty's thrombosis as an *uncaused* event, since it seems reasonable to say that Betty's taking the birth control pills merely caused the *probability of thrombosis* to be 0.2, and that the subsequent thrombosis happened by *pure chance*.⁴² If this view is correct, then we should reject the mistaken intuition, in which case there is no ambivalence to be explained (unless we take the goal of the theory to be explaining mistaken intuitions). There is consequently no need to postulate separate concepts of causation to explain any ambivalence.

But suppose for a moment that we reject this view and accept that there *is* real ambivalence. We could still resist the introduction of the net/component cause distinction by providing a plausible *alternative* explanation of the supposed ambivalence. One alternative explanation cites the indeterminate truth value of the following counterfactual as the source of uncertainty: 'If Betty hadn't taken the pills, the thrombosis would not have occurred.' If her thrombosis would have occurred anyway, one might be less inclined to judge that her pill taking was cause. But if her thrombosis would *not* have occurred then it seems clear that her pill taking *would* count as a cause. We do not know whether this counterfactual is true or false. All we know is that the *probability* of her thrombosis would have been 0.3. This is not just a question of ignorance; there *is* no determinate fact of the matter. Any semantic indeterminacy in the

⁴² Philosophers are admittedly somewhat divided on this issue. Lewis (1986), for example, thinks that it is acceptable to describe such situations as causal. A fuller discussion of this topic is provided in Hitchcock (2003, 2004).

indeterministic case is therefore also explicable by the indeterminacy regarding the truth value of the counterfactual, 'If Betty hadn't taken the pills, the thrombosis would not have occurred.' If this explanation is correct then again, there is no obviously pressing need to introduce new concepts of net and component causation.

I do not think that the concepts of net and component causation have no place in the philosophical study of causation. In a variety of applied scenarios (such as predicting the results of interventions, experimental design, constructing algorithms for causal inference, etc.) it is helpful to stipulatively define them, as they bear a clear relation to these practices. For example, a clinical trial carried out on the general population would deliver the net effect. If carried out on a population of *infertile* sexually active, non-smoking women under 35, the trial would yield the component effect. This is because the second preventative route through pregnancy would no longer be active (since none of the subjects can get pregnant), and hence only the component route would be expressed. But, I claim, these stipulative definitions do not reflect any conceptual ambiguity about the English word 'cause' and its proper use in ordinary causal discourse.

Hitchcock makes further comments relevant to the feasibility of a univocal conceptual analysis of token causation.

There are...reasons why the argument [that there is no unique causal relation] ought to be accepted as persuasive. ...[T]he concept of causation is often thought to be important to philosophy because it is an ingredient in other important concepts such as explanation, prudential rationality, and moral responsibility. What I wish to suggest is that while these concepts are indeed causal in the broad sense, the causal component of these concepts can be understood in terms of stage-one facts alone [counterfactuals, probability relations, etc.]; we do not need a unique causal relation in order to analyze these concepts. In each example we will discuss, there is a dispute as to whether one event or factor *really causes* another. Yet I invite the reader to ask herself whether anything of importance hinges upon the answer to this question. Would our understanding of why the outcome occurred be enhanced? Would we be better placed to make decisions about how we should act if we knew? Would we be better placed to assign praise and blame? The answer to each of these questions is a resounding 'no'. What possible grounds could we

have, then, for caring whether [some putative C] really caused [E]? (Hitchcock, 2003 p.9-10).

Now, one might take the above passage to be proposing the following argument: “Because we do not *need* a single concept of causation, token causation cannot be given a univocal analysis. Such an argument is obviously unsound. Let us see why. Call the concepts that have been thought to supervene on causation (e.g. explanation, moral responsibility) *supervenient*, and those on which causation has been thought to supervene (Hitchcock’s ‘stage one’ concepts) the *subvenient* concepts. The argument that no univocal analysis can be given because the supervenient concepts can be analyzed in terms of the subvenient is unsound for the following reason: Imagine that we *could* define the token causation univocally. We could then still analyze supervenient concepts in terms of subvenient stage-one facts, and consequently would not need a univocal analysis of causation for this purpose. So the fact that we can analyze the supervenient in terms of the subvenient says nothing about whether or not causation itself can be analyzed univocally.

A better reading of this passage would be to take it to be recommending the abandonment of the search for a univocal analysis *because nothing of metaphysical importance hangs on it*, rather than because it cannot be done. Given that the subvenient facts are able to do all the work that has to date been placed on the shoulders of causation, we have no reason to care whether causation can be given a univocal analysis. Now, it may well be, as Hitchcock suggests, that nothing of a fundamental *metaphysical* nature hangs on whether one event or factor *really causes* another. As I will discuss in chapter 5, our ordinary concept of causation does seem to be highly pragmatic, and our everyday causal intuitions do seem to be swayed by a variety of somewhat subjective judgments about the moral and intentional status of putative causal agents. As such, our ordinary concept of causation may not be suitable for inclusion in a systematic metaphysical picture of the world. If we do wish to give causation a sizable role in an objective metaphysics, a revisionary (rather than descriptive) analysis is clearly preferable. Alternatively, we could simply *leave out* causation and provide a metaphysics that uses only stage-one facts, which, as Hitchcock, notes, appear to be able to do all the work that causation can do. For example, we might replace the phrase “causal structure of the world” by “counterfactual structure of the world”, or “network of physical processes”. That said, however, there are still *semantic* grounds

for caring whether or not some C is a cause of E and trying to provide an analysis of the ordinary concept. The analyzability of subvenient concepts in terms of the subvenient is irrelevant to the feasibility of the semantic project.

4.4 HALL'S TWO CONCEPTS OF CAUSATION

Let us now turn to another variety of causal pluralism that is inconsistent with univocality: a pluralism at Hitchcock's stage one. Hall (2004) proposes that there are two concepts of causation: production and dependence. Hall's definition is:

(TC) C is a cause of E if and only if (E depends on C) or (C *produces* E).

Each disjunct is given a different analysis. Dependence is just counterfactual dependence. Hall does not attempt a definitive analysis of production, but says that "we evoke it when we say of an event C that it helps generate or bring about or produce another event E." Whatever production is, it is a local, intrinsic and transitive relation. Hall tentatively advances the hypothesis that the producers of E are those events that are minimally sufficient for E, in appropriate circumstances. Sober (1985) suggests that this sort of productive relation might be usefully analyzed in terms of energy-momentum transfer. One also thinks of Dowe's 'conserved quantity exchange' as a candidate for production.⁴³ I will not pursue these possibilities here.

Both disjuncts of (TC) are sufficient for C to be a cause of E. Note that dependence and production are frequently *co-instantiated*. For example, in the paradigmatically causal billiard ball collisions the motion of the second ball is both produced by the motion of the first *and* depends on it. We might dub this relation 'productive dependence.'

It is worth noting that causation, as defined by Hall is not ambiguous in the sense in which words like 'bat', 'bank' and 'pen' are ambiguous. In these cases, the two disjuncts (e.g. river bank and savings bank) are generally not co-instantiated in the same particular; their

⁴³ Dowe's conserved quantity (CQ) theory is problematic, however, since it does not apply in other worlds whose physics is different from our own. For example, (CQ) would not give the intuitively correct verdict for a billiard ball collision in a nearby world where energy-momentum is not perfectly conserved, but instead only 99% conserved. As Earman (1974) has remarked, it seems intuitively right to say that we would still judge a billiard ball collision in this nearby world to be causal even if energy-momentum was not perfectly conserved.

extensions do not overlap. Production and dependence, on the other hand, very often *are* co-instantiated in the same particular, as in the billiard ball case. (Exceptions such as omissions are sometimes viewed as rather atypical and less paradigmatic instances of causation). It is an accident that we use the same word ‘bat’ for both the nocturnal flying mammal and a piece of sports equipment. The meanings of this word are not related in any interesting way, and the two types of bat share few significant properties. Bat₁ and bat₂ are merely homonyms.⁴⁴ In the case of causation, however, the production and dependence senses do seem to be closely related, in interesting ways. For example, they are both able to play similar roles in explanation, prediction, agential control, and so on. It is no accident that the same word ‘causation’ is used for both production and dependence. Causation exhibits *polysemy* rather than *homonymy*. The senses are so closely related that only a few trained philosophers, such as Hall, would think to tease them apart.

4.4.1 Preemption, Prevention and Omission

How does (TC) fare with regard to the canonical counterexamples to the major univocal theories? I will not attempt to provide a comprehensive survey here; I will restrict myself to pointing out some of the major advantages of Hall’s dualistic theory. Recall from chapters two and three what a thorn in the side preemption has been for counterfactual analyses of causation. (TC) takes care of preemption with impressive ease. In Trainee and Supervisor, Trainee’s shot is a cause of Victim’s death because of its local, productive relationship (via Trainee’s bullet) with Victim’s death, despite the absence of dependence. It is extremely plausible that when making intuitive judgments about preemption, it is this local productive relation that we are paying attention to. This is a very natural psychological diagnosis of our intuition-forming process. Early and late preemptions count as cases of causation *not* in virtue of any dependence of the effect on the cause (while holding fixed the redundant backup) as the counterfactualist would have it; rather, C is a cause of E in virtue of the productive relation between the two.

⁴⁴ Their shared properties are rather uninformative and do not seem central to the meaning of either homonym. For example, they both share the property of being physical objects.

(TC), since it does not need to appeal to the Holding Fixed strategy in order to deliver the intuitively correct verdicts in cases of preemption, has the great advantage of not thereby ruling in Switching and Self-Canceling Threats. In Switching and Self-Canceling Threats, there is no dependence *simpliciter* between the putative cause and effect, and the latent dependencies that would be revealed by holding fixed certain facts *remain* hidden, as we desire. (TC) is thus an important advance on univocal counterfactual theories.

(TC) also deals straightforwardly with the omission and prevention counterexamples that plague the physical process theories of Salmon (1984, 1997) and Dowe (1992, 2000). My gardener's not watering my plants caused their death in virtue of the dependence relation that links the two, despite the absence of any local physical process linking the two events.⁴⁵ Cases of double prevention (where C causes E by preventing a preventer of E) are also handled successfully.⁴⁶ For example, consider an example from Hall (2004):

Bombing Mission: Suzy is piloting a plane on a mission to bomb a particular target. She is escorted on this mission by Billy in second plane. Enemy's fighter approaches, intending to shoot down Suzy's plane. However, Billy shoots first, and Enemy's plane goes down in flames. Suzy proceeds to the target and completes the mission.

Billy's action prevents Enemy from preventing Suzy's bombing. Hall claims that Billy's shooting is intuitively a cause of the bombing. Although no local process connects the two, (TC) rules that Billy's shooting is a cause since the bombing depended on it.

In addition to (TC)'s success with recalcitrant counterexamples to univocal theories, Hall provides a more general argument in favor of splitting our concept of causation in two: it enables us to preserve several of our important platitudes about causation. Locality, intrinsicness and transitivity do not apply in cases of omissions, double preventions and the like. Yet they always apply in cases of production.

⁴⁵ Strictly, the death of the plants does not depend on the gardener's not watering them. The plants, being mortal, would have died sooner or later. In order to make the counterexample work, some other detrimental effect on the plants (due to their not being watered) should be chosen as the effect. Alternatively, we could precisify the effect (e.g. the plants' death at time t).

⁴⁶ Hall's causation by double prevention is similar to Schaffer's causation by 'disconnection', where C causes E by disconnecting whatever was blocking E. See Schaffer (2000).

4.4.2 Counterexamples to (TC)

Unfortunately, there are several counterexamples to (TC). There are cases that we intuitively judge to be causation, but which exhibit neither production nor dependence. There are also cases that exhibit production and/or dependence, yet which are not causation.

Hall himself provides a counterexample to (TC), which is a slight variant on Bombing Mission:

Second Escort: Suzy is this time escorted on her mission by Billy and Mary. Again, Billy shoots down Enemy's plane and Suzy completes the mission. But this time, if Billy hadn't shot down Enemy, Mary would have. (Hall 2004).

This example is an instance of 'Preempted Prevention'; Billy's prevention of Enemy's attack preempts Mary's preventing. As in Bombing Mission, there is no production, and now, in addition, the introduction of the backup preventer Mary breaks the dependence between Billy's shooting and the bombing. Yet, if we accept Hall's intuition that the shooting is a cause of the bombing in Bombing Mission, it seems that we must make the same judgment in Second Escort.

This example illustrates an important general point. If we make just one link in a transitive causal chain non-local, there will be no productive relation between C and E. By adding in a redundant backup, we can also remove any dependence. We can thus generate counterexamples to (TC) at will, by altering familiar examples. Hall explicitly acknowledges that such counterexamples exist and leaves them as important 'unfinished business'. Here is a similar example:

Victim's Flowers: Add to the Trainee and Supervisor example the subsequent death of Victim's flowers.

Having bled to death at the hands of Trainee, Victim was unable to water his flowers. The death of the flowers does not depend on Trainee's shot (because Supervisor would have shot Victim had Trainee not done so) and there is no productive relation between these events either, since

there is no local connection between Trainee and the plants' death. Yet Trainee's shooting seems intuitively to be a cause of the plants' death.

It seems plausible that what is going on psychologically when we make our intuitive judgments in cases like these is the following: We naturally break down these cases into two discrete steps. In Bombing Mission, the first step consists of Billy's shooting down Enemy's plane. The second step consists of the omission of Enemy's attack on Suzy and the subsequent bombing. Intuitively, each of these constituent steps is clearly causal. We then implicitly chain these two causal steps together and conclude that Billy's shooting was a cause of the bombing. We can tell a similar story for Victim's Flowers. While this may seem a satisfactory account of our judgment process in these cases, it will not do as an analysis. We *know* that causation is not always transitive. Furthermore, we can generate related counterexamples in which we *cannot* appeal to this chaining strategy. This can be done by starting from an ordinary case of early preemption such as Trainee and Supervisor, and making the productive link non-local. For instance:

Action at a Distance Guns: Trainee and Supervisor are armed with action-at-a-distance guns. Trainee shoots first and Victim vaporizes. If Trainee hadn't shot, Supervisor would have, and Victim would have been vaporized in exactly the same manner.

In this case there is neither production nor dependence, yet our intuition that Trainee's shooting is the cause of Victim's vaporization remains solid. Neither can we point to a chain of intuitively causal steps. A second method of generating such counterexamples is to begin with an omission, and add in a redundant backup omission:

Patricidal Brothers: Jack and Bobby are tired of waiting to inherit their father Joe's money and independently decide to do away with him. Each day, Joe must take two pills (one red, one green) in order to keep him alive, which he receives in a rather unusual manner. Before going to bed, Jack, and then Bobby, each leave a single pill on the kitchen table. When the old man gets up in the morning, he takes the two pills. One evening, Jack, unable to wait any longer for his inheritance, decides not to leave his red pill on the table, and retires for the evening. A few minutes later, Bobby, who has decided

on the same course of action, notices that his brother Jack has not left his pill on the table. Bobby, not wanting to risk being incriminated for his father's death, leaves his green pill on the table as usual. But if Jack *had* left his red pill on the table, Bobby, wanting to ensure Joe's demise, would *not* have left his green pill. Joe awakens the next morning and, deprived of his medication, dies shortly thereafter.⁴⁷

It is perfectly clear that Jack's omission is a cause of Joe's death. Yet there is neither production nor dependence: Jack's omission merely preempted Bobby's omission. This case is interesting in that it challenges Hall's hunch that there could be nothing more to causation by omission than counterfactual dependence.⁴⁸ In virtue of what then, does Jack's omission count as a cause, if not dependence? One is initially tempted to answer "Dependence, holding fixed the fact that Bobby *did* deliver his green pill." Yet we saw in chapters two and three what havoc Holding Fixed wreaks, dragging in Switching and Self-Canceling Threats as *bona fide* types of causation. So where does this leave us? Why do we judge Jack's omission to be a cause of his father's death? I will return to this question in chapter five.

It is more difficult to construct examples in which there is both production and dependence but which are intuitively *not* cases of causation. Here is one attempt, however:

Necessary Irradiation: A stable, non-radioactive atom is irradiated, and its probability of decay goes up from zero to 0.01. The atom subsequently decays.

In this case, the atom is productively linked to the irradiating device (via the radiation), and its decay depends on the irradiation, since it would not have decayed had it not been irradiated. Yet one might argue (as above) that the irradiation did not cause the decay; it merely caused the probability of decay to be 0.01, with the subsequent decay happening by pure chance. However, as I remarked in section 4.3, not all philosophers agree with this intuition, and hence its force as a clear counterexample to (TC) is diminished.

⁴⁷ Cases with an analogous structure can be constructed in which an individual refuses to vote for a certain proposition for which unanimity is required.

⁴⁸ Hall (2004).

4.5 CONCLUSION

I do not think that the Birth Control Pills example forces us to provide a pluralistic analysis of deterministic token causation. While it is useful to define the concepts of net and component causation in relation to our epistemic practices in science, they are not required for the analysis of our ordinary concept of causation.

Hall's pluralistic theory (TC) delivers the correct verdicts for early and late preemption, and since it does not appeal to the Holding Fixed strategy, it has the great advantage over counterfactual theories (H1), (H3) and (DF) of not ruling in Switching and Self-Canceling Threats as causation. (TC) also correctly handles cases of preventions and omissions, which constitute counterexamples to physical process theories. (TC) thus constitutes a major advance on extant univocal theories of causation. There are a few counterexamples to (TC), however, and we therefore have reasons to suspect that if causation *is* a non-univocal concept, Hall's non-univocal analysis is not quite the right one.

In the next chapter I will discuss two other broadly pluralistic approaches to causation. These approaches postulate very different conceptual structures from the disjunctive form proposed in (TC) but offer hope of clearing up at least some of the remaining difficulties that the production-dependence theory faces.

5.0 THE PROTOTYPE THEORY OF CAUSATION

...you will not see something that is common to all [games], but...a complicated network of similarities overlapping and criss-crossing...I can think of no better expression to characterize these similarities than “family resemblances”.

- Ludwig Wittgenstein, *Philosophical Investigations*.

The chief danger to our philosophy, apart from laziness and woolliness, is scholasticism, the essence of which is treating what is vague as if it were precise and trying to fit it into an exact logical category.

- Frank Ramsey, *Philosophy*.

5.1 INTRODUCTION AND OVERVIEW

In the previous chapter we saw that while Hall’s disjunctive theory of token causation is able to handle *some* cases that are problematic for counterfactual theories, it is still subject to several counterexamples. In this chapter I will sketch an alternative (broadly) pluralistic theory, according to which our concept of causation has a *prototype* structure, consisting of a large cluster of properties, none of which appear to be individually necessary. In addition to doing a better job with recalcitrant counterexamples, this theory explains several features of the concept that are otherwise difficult to explain via a classical necessary and sufficient conditions analysis. It is also attractive in that it appears to be empirically testable to some degree.

In section 5.2 I put forward five theses about the nature of our ordinary concept of causation that I suggest stand in need of explanation: in particular the existence of so-called ‘prototype effects’. I also present three further desiderata that ideally a theory of causation should satisfy. In section 5.3, I discuss resemblance to paradigm theories. In section 5.4, I list twelve prototypical properties that the linguist George Lakoff thinks causation has. In section 5.5, I develop the prototype theory of Lakoff in to a cluster theory of causation, and offer some arguments in favor of this theory in section 5.6.

5.2 FIVE THESES AND THREE DESIDERATA

I will first advance five theses about the nature of the concept and how it appears to function in ordinary language.

1. Counterexamples: There are many extant univocal theories of causation and all of them have counterexamples.
2. Disagreement: There are some cases about which individuals *disagree* in their intuitive causal judgments.
3. Vagueness: There are *borderline* cases of causation.
4. Error: Individuals' intuitions are sometimes *clearly* mistaken.
5. Degrees of Typicality: Some cases of causation appear to be 'better' or more typical examples of the concept than others.

I take these five theses to be true. But what is it about the nature of our concept of causation that *makes* them true? I suggest that an adequate theory of causation should have something to say about *why* these five theses are true.

In addition to being able to explain why theses 1-5 are true, I suggest three other desiderata for an adequate theory of causation:

1. The theory should deliver the same verdicts as intuition. Ideally, when our intuitions are indeterminate, the theory should go indeterminate.⁴⁹ And when our intuition is that the case in question is a very atypical/poor example of causation, the theory should predict this.

⁴⁹ Lewis (1986) and Hiddleston (2005) make somewhat similar points

2. The theory should provide a psychologically plausible explanation of *how* we arrive at our intuitive judgments about whether one event is a cause of the other. Which criteria do we pay attention to or take into account (either explicitly or implicitly) in making those judgments? If a theory of causation can explain in a psychologically natural way, it is to be preferred (*ceteris paribus*) to a theory that, while delivering the same verdicts as intuition, does so via excessively technical and complex philosophical machinery.⁵⁰

3. It would be an advantage if a theory of causation were able to do some other philosophical or scientific work: for example, in aiding the analysis of other concepts. As Ned Hall (2004) has argued, it is not clear that a *univocal* concept of causation can satisfactorily analyze *all* of the concepts that have been taken to supervene on causation (persistence, explanation), ground causal decision theory, and so on. Hall suggests that we may need *different* concepts of causation for different philosophical purposes. Relatedly, our analysis of causation should not leave mysterious why our trusted epistemic methods for discovering causes (such as laboratory experiment, randomized clinical trials and causal inference from statistical data) are effective. If our analysis of causation seems totally unrelated to our best methods for discovering causes, this is reason to be suspicious of our analysis. Woodward (2003) and Cartwright (forthcoming) both emphasize the importance of this point.

I do not say that it is *necessary* that a theory of causation be able to explain why theses 1-5 and satisfy the additional desiderata 1-3. But, *ceteris paribus*, a theory that can do so is to be preferred over one that cannot.

Let us now look in more detail at our five theses.

1. Counterexamples: One might cynically describe the last thirty years of philosophical work on causation as a series of repeated failures to provide a univocal analysis of the concept. It has proved extremely difficult to provide necessary and sufficient conditions, and one fears that the increasingly technical apparatus that is being brought to bear on the problem may merely amount to the piling on of epicycles. In chapters two and three I provided several counterexamples to

⁵⁰ This desideratum inspired by brief remarks by Hitchcock (2001).

two of what I consider to be the most sophisticated univocal analyses of token causation. Counterexamples to the other univocal theories (nomic, probabilistic, manipulationist, physical process) are well known and I will not rehearse them here.⁵¹ Several questions arise: First, why is it that there have been *so many* different kinds of theories of causation, and why have they all seemed plausible (at least initially) to some philosophers. Why is it that each theory appears to be able deliver the correct verdict for many cases (including problematic ones) but falls frustratingly short of universality? Can we explain the successes and failures of the various theories of causation, and why those theories seemed attractive in the first place. Is there something about the nature of our concept of causation that makes it *inevitable* that univocal theories have the patterns of successes and failures that they do?

2 and 3. Disagreement and Vagueness

Consider the following case:

Automobile Accident: You are driving steadily down a deserted highway when suddenly, without warning, a truck ploughs into the side of your car. It is later revealed that the driver of the truck was heavily intoxicated and had run a red light.

Was you driving down the highway a cause of the accident? One instinctively replies ‘no.’ But recall that the accident would not have happened had you not been driving down the highway. If we change the example and replace the car and the truck by billiard balls, we would say that each ball’s motion was a cause of the collision. Does this second example merely indicate that we have conflated causation with moral responsibility? Perhaps. But why not take this example to indicate that (human) causation is in part a moral concept?

Consider also the following case of symmetric overdetermination:

Gas Pellets: Billy and Suzy simultaneously push down with their right index fingers on a button that releases gas pellets into an execution chamber. Mary, the condemned prisoner, is asphyxiated. Billy and Suzy each push with sufficient force so that if the other hadn’t pushed, the button would still have depressed.

⁵¹ But see Lewis (1973), Woodward (2003), Schaffer (2000, 2005)

Is Billy's action a cause of Mary's asphyxiation? In order to preempt one possible objection, let us further state that Mary would have died exactly the same death, at exactly the same time, had Billy not pushed. (We may assume, for instance, that Billy and Suzy push at 11:55pm, and that the pellets are programmed to drop at midnight, the scheduled time of execution, as long as the button has been depressed by that time). Then we cannot say that Billy's action *was* a cause of the actual specific death that occurred by arguing that if he had not pushed, Mary would have died a slightly different death. Intuitions are clearer about other cases of symmetric overdetermination, such as:

Forest Fire: Cigarette A is dropped on the west side of a large forest; cigarette B on the east. Two fires start, spread towards the center of the forest and eventually merge, incinerating the whole forest. Had only Cigarette B been dropped, the forest would still have been completely incinerated.

In this case, it seems entirely unproblematic to say that Cigarette A was a cause of the forest fire. But Gas Pellets is far less clear. Cases of symmetric overdetermination should not all be considered alike; the details of the particular processes matter. While Mary's asphyxiation in no way depends on Billy's action, it seems that in some way he *is* causally responsible for what happened. Not also that if the button pushing merely released a small trap door that allowed the pellets to fall, there is no transfer of energy-momentum from Billy to Mary. Let us muddy the waters even further by replacing Suzy by a robotic hammer that depresses the button automatically, rather violently, at 11:55pm. As the hammer makes contact with the button, Billy lightly pushes on the hammer. Now it is perhaps even less clear whether or not Billy is a cause.

Next consider the following *series* of cases, the first of which was discussed in chapter two:

Flip: A trolley is hurtling down a track towards Victim. The track diverges into two 100-yard subtracks, which then reconverge. Victim is strapped to the track just beyond the reconvergence point. Just as the trolley is approaching the divergence point, Suzy flips a switch (F) that takes to trolley onto the left subtrack (L). The trolley travels the 100 yards of this subtrack, and then regains the main track, crushing Victim (V). Had Suzy not

flipped, the trolley would have taken the right subtrack (R), but Victim would have nonetheless been crushed.

Flip Shove: Just as the trolley is approaching the divergence point, Suzy pushes the trolley from behind with just enough force to make it derail. The trolley continues, cross-country, but subsequently collides with the main track at the exact location where Victim is helplessly strapped. Victim is crushed. Had Suzy not pushed, the trolley would have taken the right subtrack (R) and Victim would still have been crushed.

Flip Bulldozer: Just as the trolley is approaching the divergence point, Suzy drives her bulldozer into the trolley at eighty miles an hour, such that it derails. Suzy powers forward cross-country and bulldozes the trolley right into Victim, who is crushed. Had Suzy not bulldozed the trolley, the trolley would have taken the right subtrack (R), and Victim would still have been crushed.

Our intuitions are clear, I take it, for Flip and Flip Bulldozer. In the former case, Suzy's flipping is *not* a cause of Victim's crushing; in the latter, Suzy's bulldozing *is*. But what about Flip Shove? This series of cases illustrates what I call "The Switching to Preemption Transition". Switching is generally not thought of as causation; preemption generally *is*. One should not be surprised to find borderline cases at the transition.

Finally, consider:

Irradiation: A particle has a 0.01 chance of decaying in time t . It is irradiated and the chance of decay goes up to 0.99. (These probabilities represent genuine, irreducible indeterminism). The particle then decays.

Is the irradiation a cause the decay?⁵² On the one hand, the fact that the irradiation greatly increased the probability seems to incline one towards thinking that the irradiation caused the decay. On the other hand, it might be argued that that the irradiation merely caused the

⁵² This case is discussed by Hitchcock (2003).

probability of decay to change, which it then did, but for no reason, i.e. the decay was uncaused. In addition, there appears to be no determinate fact of the matter as to whether the particle would have decayed had it not been irradiated – although it is very likely that it would *not* have.

If Irradiation seems clear cut, consider:

Irradiation*: A particle has a 0.98 chance of decaying in time t . It is irradiated and the chance of decay goes up to 0.99. The particle then decays.

Is the irradiation a cause of the decay? There is still an increase in probability due to the irradiation, but to my mind at least, it is much less clear whether the irradiation can be said to be a cause of the decay. In this case it is highly probable that the particle *would have* decayed even if it had *not* been irradiated. Suppose the irradiation merely changes the probability of decay within the next 24 hours, and 22 hours later the particle decays. Does this time lag alter our intuitions?

I have no systematic empirical data on what people's intuitions actually are regarding these cases; to my knowledge, there is no systematic data on *any* of the canonical examples in the causation literature. But in my experience, the above examples (Automobile Accident, Gas Pellets, Flip Shove and some variants of Irradiation) elicit a fair amount of disagreement and uncertainty. One individual (S1) may feel that C is a cause of E, while S2 feels that C is not a cause of E. S3 may feel simply unable to make a determinate judgment. I take this uncertainty and disagreement to be an interesting feature of the concept of causation, which stands in need of explanation. Note that the vagueness in the four cases above does not appear to be of the same *quantitative* nature as that which generates *sorites* series (with regard to baldness for example). It seems closer in nature to that associated with concepts such as religion and friendship, where we are not clear what factors must be present, or which combinations of factors must be present for some practice to count as religion, or for two individuals to stand in the 'friendship' relation. I shall return to this point later in this chapter.

4. Error

Individuals' intuitions are sometimes *clearly* mistaken. I argued in chapter two that in the Two Assassins and Avalanche Warning examples, the intuition that the yells are causes of the survival

and subsequent good health are mistaken, and provided reasons for thinking so. In these cases, individuals will *retract* their original judgment when reminded of particular facts of the case and agree that they were mistaken. It is a common phenomenon that individuals frequently confuse correlation with causation. If told that children who drank large amounts of diet cola sweetened with saccharin in the 1980s now suffer a significantly greater incidence of stomach cancer, many will assert that diet cola causes cancer.⁵³ But they can be easily made to admit their error when presented with the correlation between shoe size and reading ability in children and asked whether this correlation is causal.

While we do not want a theory of causation to deliver verdicts that accord with *mistaken* intuitions, I claim that widespread systematic errors stand in need of theoretical explanation. It counts in a theory's favor if it can also provide a *Theory of Error*.

5. Degrees of Typicality

Some cases of causation appear to be 'better' or more typical examples of the concept than others. A billiard ball collision is perhaps considered to be the paradigmatic case of causation. Other varieties of causation seem less 'typical'. 'Causation' by omission has generated a fair amount of controversy. While some (e.g. Dowe 2000, Beebe 2004) do not consider it to be causation at all, there appears to be a general consensus in the causation literature that omissions *can* be causes. But most would admit that the mother's momentary inattention is a *less typical* example of a cause of her child's death than is the truck that ran into her child. In Automobile Accident above, even if one admits that driving along the highway is a cause of the accident, it is in a less 'central' sense than that in which the intoxicated truck driver is a cause.

The fact that certain instances of some concepts are judged to be more representative examples than others has been well-confirmed empirically. The pioneering work of Rosch and associates in the 1970s demonstrated that many concepts display such 'prototypical effects'. Participants in these experiments were:

1. Asked to rate (on a scale from 1-7) how good an example of the category (e.g. bird) various members are (e.g. robin, chicken, etc.)
2. Tested for the speed of their reactions to question such as, "A chicken is a bird. True or False?"

⁵³ Saccharin is not now considered a carcinogen in humans.

3. Asked to produce examples of the category (by listing or drawing).

Each of the above experiments indicated that certain members of a category are indeed judged to be more typical than others. In addition, an asymmetry in similarity ratings was detected: less representative examples are considered to be more similar to more representative examples than *vice versa*. The degree of family resemblance (as measured by ‘perceived similarity’) correlated well with numerical ratings of the examples (see Lakoff 1987, p.39-46 for further discussion).

I have seen no such empirical work on causation, although Lakoff (1987) claims that such prototype effects exist for causation. It seems exceedingly likely that such prototypical effects *would* be confirmed empirically and it is somewhat surprising that philosophical literature is so completely lacking in this respect.

The classical theory of concepts, according to which all instances of a concept share some common ‘essence’ that can be expressed in terms of necessary and sufficient conditions has no resources to explain prototype effects. Rosch saw her work in Wittgensteinian terms, with instances of a category related to one another by ‘family resemblance’ rather than sharing some essence.

How are Rosch’s prototype effects to be interpreted? To what extent do they constrain the structure of our concepts? And how might these considerations apply to causation? Could some prototype theory explain why the other four theses are true and satisfy our four desiderata? I suggest that the answer to this question is ‘yes’.

5.3 RESEMBLANCE TO PARADIGM THEORIES

Shortly after the publication of Wittgenstein’s *Philosophical Investigations*, in many areas of philosophy, the possibility of providing classical definitions of concepts was questioned. In aesthetics, for example, there was a lively debate over whether art could be defined.⁵⁴ Family resemblance accounts were proposed, according to which there is nothing that all artworks share. Instead, artworks are judged to be ‘art’ in virtue of their family resemblance to some paradigmatic case (such as the *Mona Lisa*). Such accounts were fairly uniformly rejected because it was not clear in virtue of what something counted as paradigm case, and the notion of

⁵⁴ See, for example Weitz 1956, Mandelbaum 1965.

resemblance seemed empty: what did it consist in, and how much was enough? Lacking answers to these questions, resemblance to paradigm accounts fell out of favor.

Anscombe (1971) put forward a theory of causation that can be couched in Wittgensteinian terms. She writes:

“The word ‘cause’ is itself highly general...I mean: the word ‘cause’ can be *added* to a language in which are already represented many causal concepts. A small selection: scrape, push, wet, carry, eat, burn, knock over, keep off, squash, make (e.g. noises, paper boats), hurt. But if we care to imagine languages in which no special causal concepts are represented, then no description of the use of a word in such languages will be able to present it as meaning cause.” Anscombe (1979, p.93).

The key idea here is that the concepts ‘scrape’ and ‘burn’ are semantically prior to ‘cause’; the generic ‘cause’ is an abstraction from these *specific* causal concepts, and is parasitic on them. It does not, however, exhaust their content.⁵⁵ Godfrey-Smith (forthcoming) has recently outlined an Anscombian theory of causation:

Let S be a set of causal verbs and other linguistic formulas which represent “special causal concepts” in Anscombe’s sense. Then C is a cause of E if and only if the relation between C and E can also be described using some member of S. (Godfrey-Smith, forthcoming).⁵⁶

⁵⁵ Nancy Cartwright (1999, 2004, forthcoming) has also advanced a similar theory of causation. Note that such an account is pluralistic to an *extreme* degree. Since the generic ‘cause’ doesn’t exhaust the content of each specific causal verb, each causal verb represents a different kind of causation.

“The term ‘cause’ is highly unspecific. It commits us to nothing about the kind of causality involved.” If, as I claim, there is no such thing as *the* causal relation, what are we to make of claims of the form ‘X causes Y’?... The term ‘cause’ is abstract...Whenever it is true that ‘X causes Y’, there will always be some further more concrete description that the causing consists in. This makes claims with the term ‘cause’ in them *unspecific*. The cat causes the milk to disappear; it *laps* it up. Bombarding the population of atoms of a ruby-rod with light from an intense flash causes an inversion of the population; it *pumps* the population to an inverted state.” Cartwright (1999, p.119-120).⁵⁵

Cartwright calls general, abstract, unspecific causal claims “thin”, and the more specific causal verbs “thick” causal concepts, following Bernard Williams’ contrast between terms like *good* and *ought* with “‘thicker’ or more specific ethical notions ...such as *treachery* and *promise* and *brutality* and *courage* (Williams, 1985, p.129). In another paper she states that “All thick causal concepts imply cause. *Compressing* implies *causing* + x.” (Cartwright 2004, p. 817).

Hitchcock (forthcoming) questions whether this unexhausted content is really *causal* in nature.

⁵⁶ As it stands, such a theory is implausible. As Lakoff (1987) has argued, we tend to use the generic ‘cause’ only in cases in which there is *no* specific causal verb that could be substituted. Think, for example, of the surgeon’s error causing an individual’s inability to work and subsequent loss of home. What specific causal verb could be substituted here? Godfrey-Smith recognizes this shortcoming and suggests that the causal relation can be described

He adds that *new* cases of causation may be added if they are judged to display sufficient family resemblance to members of S.

But such a theory faces the same objections that resemblance to paradigm theories of art faced. First, what determines membership of S? A dilemma presents itself: If *no* general criteria can be provided, why take seriously such an account? But if we *can* give general abstract criteria (in terms of the properties that determine membership of S, why not simply provide a theory in terms of those properties?

Second, what does family resemblance consist in and how much resemblance to existing member of S is enough to grant membership? It is well-known that measurements of overall similarity are problematic. Everything resembles everything else to *some* degree. To be meaningful, such measurements must be made *relative to specific properties or features*.

5.4 LAKOFF'S TWELVE PROTOTYPICAL PROPERTIES

The linguist George Lakoff (1987) claims that the prototype effects causation purportedly shows can be explained by a cluster of properties that

...seems to define a prototypical causation, and non-prototypical varieties of causation seem to be best characterizable in terms of deviations from that cluster. Prototypical causation appears to be direct manipulation, which is characterized most typically by the following cluster of properties:

1. There is an agent that does something.
2. There is a patient that undergoes a change to a new state.
3. Properties 1 and 2 constitute a single event; they overlap in time and space; the agent comes in contact with the patient.
4. Part of what the agent does (either the motion or the exercise of will) precedes the change in the patient.

using some concatenated *chain* of specific causal verbs. But what then of omission? What specific causal verb could be substituted for 'cause' in "The mother's inattention was a cause of her child's death (under the wheels of a truck)"?

5. The agent is the energy source; the patient is the energy goal; there is a transfer of energy from agent to patient.
6. There is a single definite agent and a single definite patient.
7. The agent is human.
8. The agent wills his action.
9. The agent is in control of his action
10. The agent bears primary responsibility for both his action and the change
11. The agent uses his hands, body or some instrument.
12. The agent is looking at the patient, the change in the patient is perceptible, and the agent perceives the change.

The most representative examples of humanly relevant causation have all [twelve]... of these properties. This is the case in the most typical kinds of causation in the linguistics literature: Max Broke the window, Brutus killed Caesar, etc. Billiard-ball causation, of the kind most discussed in the natural sciences, has properties 1 through 6.” (Lakoff 1987. p.54-55).

Such a characterization takes care of two objections to the resemblance to paradigm case theories. First, the identity of the paradigm or ‘prototypical’ case is provided, in terms of properties. Second, family resemblance between two instances of causation consists in *sharing properties of this cluster*. And degree of typicality will be related to the number of properties that a specific causal relation instantiates. But questions are left unanswered by Lakoff’s account. *How many* of these properties are required? Possession of *which* subsets of properties is sufficient for describing a relation as causal? Does the absence of certain properties lead to borderline cases? If so, which? Lakoff’s list of properties may seem unusual (if not bizarre) to those who are familiar with the philosophical literature on causation. While Lakoff makes reference to properties that will be familiar to the analyst of causation such as agency, energy transference and spatiotemporal locality, he makes no reference to counterfactual dependence (which is surely of major relevance to causation), lawlike regularity, probability, and so on. In the next section, I will try to address some of these worries.

5.5. CAUSATION AS A CLUSTER CONCEPT⁵⁷

I propose a more fleshed out version of Lakoff's prototype theory, according to which causation is a 'cluster concept'. While cluster theories have been offered for a range of concepts such as art, love, race, gender, time and free will, no one to my knowledge, has presented a well worked-out cluster theory of causation. In this section I offer a sketch of what such a theory might look like. In subsequent sections I provide some arguments in favor of the cluster theory of causation.

Let us first direct our attention toward clarifying what a cluster concept *is*. I take the core ideas of the notion of a cluster concept to be close to those expressed by Wittgenstein regarding 'family resemblance' concepts:

- (1*) Cluster concepts have no 'essence' that is shared by all instances.
- (2*) Individuals falling under the concept display a family resemblance.
- (3*) The extension of the concept is vague.

On the lack of essences, Wittgenstein writes:

[S]omeone might object against me: "You take the easy way out! You talk about all sorts of language-games, but have nowhere said what the essence of a language-game, and hence of language, is: what is common to all these activities, and what makes them into language or parts of language...". And this is true.—Instead of producing something common to all that we call language, I am saying that these phenomena have no one thing in common which makes us use the same word for all—but that they are *related* to one another in many different ways...(*PI* §65).

On family resemblance:

Consider for example the proceedings that we call "games". I mean board-games, card-games, ball-games, Olympic games, and so on. What is common to them all?—Don't say: "There *must* be something in common, or they would not be called 'games'"—but

⁵⁷ This section owes much to Gaut (2000).

look and see whether there is anything common to all.—For if you look at them you will not see something that is common to all, but...a complicated network of similarities overlapping and criss-crossing...I can think of no better expression to characterize these similarities than “family resemblances” (*PI* §66-7).

And on the vagueness of extension of family resemblance concepts:

If someone were to draw a sharp boundary, I could not acknowledge it as the one that I too always wanted to draw...For I did not want to draw one at all...[I]magine having to sketch a sharply defined picture corresponding to a blurred one...won't it become a hopeless task?—And this is the position you are in if you look for definitions corresponding to our concepts... (*PI* §76-77).

Wittgenstein views the obsessive search for definitions by the conceptual analyst as a misguided and fruitless pursuit akin to a fly buzzing around inside a bottle. Wittgenstein views himself as offering a way out of this unproductive practice:

What is your aim in philosophy?—To shew the fly the way out of the fly-bottle. (*PI* §309).

Gaut (2000) offers a definition of ‘cluster concept’ which seems to capture Wittgenstein’s three core ideas:

(CL) There are a number of properties⁵⁸ or criteria that are relevant to, or ‘count towards’ an individual’s being an instance of the concept. X is a cluster concept if and only if the following four conditions are jointly satisfied:

- (1) The presence of the entire set of properties (the cluster set) is sufficient for the concept to be applied.
- (2) Some subsets are also sufficient for its application. (There may be a good deal of indeterminacy over which subsets are sufficient).

⁵⁸ Property’ is used in an informal sense here, more or less interchangeably with ‘criterion’ or ‘condition’.

(3) No property is necessary

(4) At least one property from the cluster set must be instantiated.

(1*) follows from condition (3); hence no univocal theory will work. (W2) will be true to the extent that subsets share properties. (W3) is related to (2). Note that (2) in fact follows from (1) and (3). We may think of the whole cluster set as an abstract prototype (that need not be instantiated by any actual case).

The key difference between a univocal theory and a cluster theory is that in the case of the latter, a *variety* of factors are taken into account in delivering a verdict about a particular case, rather than just a single necessary and sufficient condition.

So much for the logical form of the cluster theory. What about its content? Which properties are to be included in the cluster for causation? I am happy to accept most or all of Lakoff's twelve. But in addition, I would add counterfactual dependence, probability raising, manipulability of the effect by the cause, lawlike correlation and perhaps others. From the discussion of Patricidal Brothers (see chapter 4) in which there is neither dependence nor production, the most plausible account of our judgment would seem to be that the effect does depend on the cause when we *hold fixed* the fact that the second brother does not fail to leave his pill. So perhaps we also need to include dependence *modulo* G in the cluster set. This will not lead to the generation of Switching and Self-Cancelled Threats counterexamples, however, since we are not claiming that dependence *modulo* G is sufficient for causation, but just one property of the cluster set. To explain Second Escort (discussed in chapter four), one might include the 'property' 'C causes D and D causes E'. This does not commit us to the claim that causation is *always* transitive: nor is this circular. If C is clearly judged to be a cause of D on the basis of other cluster properties, and D the cause of E, then this will 'influence' our judgment towards thinking that C is a cause of E.

I also include moral responsibility as a cluster property in order to explain our intuition in Automobile Accident. I am not alone in making this claim. Psychologists have found empirically that our causal judgments are influenced by moral facts (Alicke 1992, Knobe MS). For certain cases of omission, Beebe (2004) has argued convincingly that moral responsibility plays a role in our causal judgments. Even Hitchcock, who has made significant contributions to the univocal analysis of causation purely in terms of counterfactuals, has recently written:

My suggestion is that the concept of token causation seems to be involved in our *post hoc* assessments of responsibility. It plays a role in our assessments of legal and moral responsibility, but also in assessments of more abstract forms of responsibility. For example, in fault analysis in engineering, or in performing an autopsy, one is trying to discover which component of a complex system is responsible for the failure of that system to function; this type of responsibility is not literally moral, but is broadly normative in character. Given this role, it is not altogether surprising that our judgments of token causation are influenced by normative considerations. If this account of the nature of token causation is correct, it is not surprising that no attempt to analyze token causation has been fully successful. The best that is to be hoped for is a piecemeal analysis of the various factors that then to pull our judgment in one direction or another (Hitchcock MSb).

Prototypical cases of causation will exhibit *all* of these properties, though none of them is *necessary* for some event C to be a cause of E. From (2) above, several of subsets of these properties are sufficient for a relation to be causal. Subsets correspond to slightly different, but closely related, senses of causation. Given the very large number of possible subsets/senses, cluster concepts are potentially highly pluralistic. However, rather than saying that we have a *plurality of concepts* of causation, I prefer to say that we have a single *polysemous* concept.

I still have not answered the question concerning exactly *which* subsets of the cluster set are sufficient for causation, and under what conditions borderline cases arise. This is a very difficult question to answer in a fully satisfactory manner and I will not attempt to give a *full* answer here. I will try, however, to make a few helpful preliminary suggestions:

1. Counterfactual dependence of E on C seems quite close, even on its own, to being a sufficient subset. Birth and Death was proposed in chapter two as a possible counterexample to its sufficiency. Perhaps some version of the distality objection could rule this case out. Some variants of Irradiation, in which the probability of decay is increased from zero to some low value, are also counterexamples, but this is arguable.
2. Hall's 'production', or a variant thereof involving some kind of energy/momentum transfer also appears to be very close to sufficiency. There are perhaps cases a few case like Flip Shove in which it is not *clearly* sufficient.

3. Neither dependence nor production are necessary (as omission and preemption counterexamples demonstrate). There are cases of causation (Patricidal Brothers and Second Escort) that lack *both* production and dependence, so there must be sufficient subsets of the cluster properties that include *neither* property. Perhaps these subsets must include the presence of a *chain* of causes linking C to E, and/or dependence *modulo* some fact G of E on C.
4. Borderline cases may arise at the Switching to Preemption Transition. The more the putative cause C contributes in the way of production (perhaps in terms of energy-momentum transfer), the clearer it becomes that C is a cause of E. The Flip-Flip Shove-Flip Bulldozer series illustrates this transition.
5. Borderline cases may also arise in *indeterministic* examples. In Irradiation variants, the semantic indeterminacy may arise from uncertainty over whether or not the condition ‘The putative cause (and background conditions) must *suffice* for the effect’ is necessary.

I said at the beginning of this section that no one to my knowledge had put forward a cluster concept theory of causation. That is not entirely true. Both Skyrms (1984) and Healey (1994) have used the term ‘cluster concept’. However, their idea of what a cluster concept is appears to differ greatly from that explicated above. Skyrms states that

“...our ordinary, everyday conception of causation is an amiably confused jumble of [positive statistical relevance, transfer of energy-momentum and manipulability], with the principles of locality of causation and temporal priority of cause to effect and perhaps a few other things thrown in for good measure...[I]n the noisy macroworld of everyday life they often go together....There is no real point in drawing ...distinctions [between statistical causation, causation as a transfer of energy-momentum and manipulability causation]...in such an environment.... The game has been altered by the progress of science. The notion of temporal priority was qualified by relativity; that of constant conjunction made obsolescent by quantum mechanics. When we ask whether causes operate locally in the quantum domain, the old cluster concept loses its heuristic value and becomes positively misleading (1984, p.254).

The Einstein-Podolsky-Rosen (EPR) thought experiment is given as an example of how science has supposedly challenged our “old cluster concept”: The relationship between the spin states of separated particles in the EPR experiment displays some of the features Skyrms mentions but not others. The spin state S1 of one of the particles raises the probability that the other particle will be in spin state S2, S2 depends counterfactually on S1, and S1 and S2 are linked by a lawlike generalization. However, there is no way to manipulate S2 via S1, and there is no local transfer of energy-momentum between the two particles.⁵⁹ Woodward (2003) interprets Skyrms as saying that there is *no fact of the matter* about whether spin state S1 causes spin state S2. Healey (1994) is more explicit:

Our concept of causation is a cluster concept which we apply on the basis of many different criteria, including the following. Causes precede their effects rather than succeeding them. Causes are, in principle if not in practice, manipulable in order to bring about their effects. If a connection between two events is capable of transmitting a mark, then it connects cause to effect. A cause and effect are linked by the transfer of some conserved quantity such as energy-momentum or angular momentum. Causes are preceding events linked to succeeding events by basic laws, appeal to which thereby explains the occurrence of the effect. Causes raise the probability of their effects. In paradigm cases, these criteria count the same pairs of events as cause/effect pairs. Moreover, there are very many such paradigm cases. But there is no *a priori* guarantee that the criteria should always deliver the same verdict. If they fail to do so for some pair of events, then some reasons will incline one to accept that this pair are related as cause and effect, and other reasons will incline one to deny this. No further appeal to the basic concept of causation will be possible – that concept will have broken down. We face just such a breakdown of the concept in the situation described [EPR]⁶⁰ ... [I]f the concept of causation has itself broken down here, how can we possibly answer the question as to whether [the relation between the spin states] is causal?...The way forward is to pick up

⁵⁹ This summary of the results of the EPR experiment is adapted from Woodward’s discussion (2003).

⁶⁰ Note also that, as Woodward (2003) points out, it is not just in unfamiliar areas of modern physics that some of the criteria fail to be satisfied; the ordinary cluster concept comes apart in more familiar cases of prevention and omission. Yet we often unproblematically judge these cases to be instances of causation.

the fragments of our shattered causal concepts and to craft different fragmentary concepts of causation for different purposes (Healey, 1994, p.371-2).

Skyrms and Healey's notion of a cluster concept seems to simply be:

1. The cluster set of all the criteria is sufficient for the application of the concept
2. There are *no* clearly sufficient subsets.

If only some of these criteria are satisfied, then, according to Skyrms and Healey, our intuitive verdict will be indeterminate.

According to this characterization, however, it appears that a cluster concept is really nothing more than a univocal concept, with necessary and sufficient conditions for its (clear) application. The cluster properties are individually necessary for C to be a cause of E, because if any one is absent the concept 'breaks down' and we cannot state that "C is a cause of E" is true.

5.6 IN FAVOR OF CLUSTER CONCEPT

In this section I will provide a preliminary evaluation of the cluster theory of causation relative to the six theses and three desiderata listed in section 5.2. Let us take the three desiderata first. Ideally our theory should:

(1) *Deliver the same verdicts as intuition.*

I will not aim to demonstrate that the cluster theory is able to deliver the intuitively correct results for all of the canonical counterexamples to other rival theories. Let me draw attention, however, to two major classes of counterexamples faced by univocal theories of causation: causation by preemption and causation by omission. Cases of causation by preemption have probably been the single greatest obstacle faced by counterfactual theories of causation. Recall Billy and Suzy: while there is no counterfactual dependence between Suzy's throw and the

bottle's shattering, there *is* a local transfer of the momentum from Suzy's arm to Suzy's rock to the bottle, resulting in its shattering. The lack of dependence does not rule out Suzy's throw being a cause. Counterfactual dependence of E on C is only one of the criteria that counts towards a relation's being causal; it is not a necessary condition.

Turn now to omission. Causation by omission has been perhaps the greatest obstacle to transference/physical process theories of causation. Consider

Gardener: My plants died when I was away on vacation. If my gardener had watered them, as he was supposed to have done, they would not have died.

It seems correct in this case to say that the gardener's failure to water the plants was a cause of their death – perhaps even *the* cause. Yet there is no obvious physical transference from the gardener to the plants; we may assume that the Gardener was never remotely close to the plants without altering our intuition. Gardener is therefore a counterexample to univocal theories based on transference. Again, the cluster theory would seem to be able to deal with such cases. We may hypothesize that transference is not a necessary criterion, but merely one of many criteria that are relevant to whether or not a relation is one of causation. The absence of physical transference does not therefore mean that the Gardener-plants relation is not causal. Note that the counterfactual dependence between Gardener's omission and the plants death appears to carry particular weight in this example. Perhaps this dependence in combination with the moral culpability of the Gardener is sufficient to judge him to be a cause.

(2) *Give a plausible explanation of how intuitive judgments are arrived at. Which criteria are taken into account (explicitly or implicitly) in making those judgments.*

The above solution to the preemption problem is extremely simple and for that reason offers a highly plausible account of what our intuitive judgment processes actually involve. In other words, it is a plausible model of the factors we take into account when making, for example, the very rapid judgment that Suzy's throw, but not Billy's, is a cause of the bottle's shattering. In order to cope with various kinds of preemption, counterfactual theories of causation have had to

propose additional, sometimes highly technical conditions. While such epicyclic additions may sometimes deliver the right results, they are often so complex as to rule themselves out as realistic models of our intuition-forming processes. Similarly, the appeal to dependence in Gardener is a plausible model of how we judge that omissions can sometimes be causes.

(3) *Do some useful further philosophical or scientific work.*

Conceiving of causation as a cluster concept helps us to distinguish different senses of causation. One sense may be useful for some purposes but not for others. For example, Ned Hall (2004) has suggested that counterfactual dependence is the sense of causation that should be used to underwrite causal decision theory. Dependence, however, is not a suitable basis on which to rest an analysis of persistence, Hall claims. Instead, some local ‘productive’ sense of causation (perhaps along physical process/transference lines, though Hall does not advocate this) is required. Dependence does not suffice. By distinguishing between these various different senses of causation, the cluster theory is able to do further philosophical work that a univocal theory is *not* able to carry out as successfully. A counterfactual theory is useful for causal decision theory, but not for analyzing persistence. The cluster theory provides the resources for tackling *all* of these tasks.

Is the prototype/cluster theory able to explain why the five theses Counterexamples, Disagreement, Vagueness, Error and Degrees of Typicality are all true?

(1) Counterexamples: *Explain why there are so many theories of causation, explain the successes and failures those theories, and why they seemed attractive in the first place.*

Univocal theories inflate what, according to the cluster theory, are merely *single properties of the cluster set* into necessary and sufficient conditions. For this reason, univocal theories fail. For example, counterfactual dependence is not necessary, as cases of preemption amply demonstrate. However, in cases where there is no preempted backup (e.g. remove Billy from Billy and Suzy), a simple dependence account does give the right result. Dependence is a feature that frequently

co-occurs with causation, and therefore counterfactual theories *do* deliver the intuitively correct verdicts in prototypical cases. Furthermore, it seems highly likely that intuitive judgments really *are* based on dependence in cases of omission. For these reasons, counterfactual dependence is an attractive starting point for an analysis of causation. But such an analysis falls short whenever there is a redundant backup.

Similarly, transference/process theories work well for billiard ball-type cases, and are attractive for this reason. In addition, in cases of preemption such as Billy and Suzy, it seems very plausible that our intuitive judgments are strongly influenced by local momentum transfer. So it is understandable that physical process theories of one form or another have been thought promising by some philosophers. But again, while physical transference commonly co-occurs with causation, there are plenty of intuitively clear cases where it does not, such as Gardener.

Both counterfactual theories and transference theories appear to get something right about causation. One advantage of the cluster account is that it is able to retain the advantages of both theories. It does not ‘throw baby out with bathwater’ by rejecting counterfactual theories in the face of preemption counterexamples. Instead, in contrast to the all-or-nothing approach of the univocalist, the cluster theory *retains* dependence as one criterial property within the cluster set; dependence contributes *something*, but not everything, to our understanding of causation.

What do univocal theories say about the successes and failures of rival theories? Nothing much, it would seem, other than that they failed because they did not start from the right criterion.

The cluster theory is able to explain why there are *so many* extant analyses. There are properties in the cluster set that are prototypically present when some C causes E. For many of these properties (regularity, dependence, probability raising, manipulability, physical transference, etc.) a corresponding univocal analysis has been constructed around it. Each theory gets *many* examples right and thus has some credibility as a candidate analysis of causation. Ultimately, these analyses are *hasty generalizations*. Attempts to fix them by adding epicycles have led to a profusion of sub-theories. Again, extant univocal theories do not appear to have anything to say about why there should be so many rival theories.

(2) and (3) *Explain Disagreements and Vagueness in intuitions.*

Indeterminacy and/or disagreement arises because there is uncertainty regarding *exactly which sets of properties* or criteria are sufficient for the application of the concept. In prototypical cases, *all* of the properties of the cluster set occur together. Borderline cases show considerable family resemblance to the paradigm case, in virtue of their shared properties, though not enough to persuade us that they are clear-cut cases; they seem ‘lacking’ in some troubling respect. This may engender a sort of ‘cognitive dissonance’ that the individual may resolve either by making a determinate true or false judgment on whether or not some phenomenon is genuinely causal. If two individuals resolve this dissonance in opposite directions, disagreement will result. Alternatively, the individual may choose to leave this dissonance unresolved and claim that the case is indeterminate, perhaps having no truth value, as we have already seen.

The putative cause in Automobile Accident lacks any moral responsibility for the accident. In Flip Shove, it is not clear whether the Shove contributes enough energy-momentum transfer to count as a cause. In Irradiation, the example lacks the property of having the cause *determine* its effect, given the background conditions.

(4) *Error: Explain why mistaken intuitive judgments are sometimes made.*

Sometimes it is possible to establish in a non-question-begging manner that certain intuitions are mistaken. While it is not generally admissible to rule that some intuition is mistaken simply because it disagrees with some or other preferred theory of causation, we may legitimately rule an intuition mistaken if it clearly conflicts with intuitions about closely analogous cases. For example, the causation-correlation confusion is a well-known informal fallacy. However, on the application of mild amounts of corrective pressure, individuals will readily admit that the inferences on which their earlier intuitions were not well-founded.

The cluster theory is able to offer a simple, and by now familiar, explanation of why intuitions are sometimes mistaken in this manner. It is easy to be tempted by the frequent co-occurrence of one cluster property with causation into inflating this single property into a sufficient condition. Some features of particular cases are particularly dramatic and have a

distracting influence on our judgments. For example, in Avalanche Warning (chapter two) one is immediately struck by the fact that the Mountain Rescue officer appears to deliver a *warning*. The fact that he therefore appears to deserve some moral credit for the survival of party of skiers (which is only one cluster property) clouds our judgment. I leave it to the reader to imagine how similarly mistaken intuitions could easily arise.

(5) *Explains Degrees of Typicality.*

This has been discussed at length above. More typical/central cases display more of the cluster set of properties. Less typical cases display fewer.

5.7 CONCLUSION

It has been my intention in this chapter to sketch an alternative, broadly pluralistic theory of causation: the cluster account. Such a theory offers greater flexibility in dealing with recalcitrant counterexamples, has resources for explaining disagreements and vagueness and is able to give an account of how prototype effects arise. Though a lot more work would have to be done on the content of this theory (principally in providing tighter constraints on which subsets are sufficient) its logical form is sufficiently promising to merit being taken seriously as an alternative to the major extant theories (both univocal and disjunctive).

6.0 CONCLUSIONS

I feel that the most significant contribution of this dissertation is the demonstration that even (what I consider to be) the most sophisticated extant accounts of token causation (the counterfactual theories that rely on Holding Fixed) face *major* difficulties. Furthermore, there do not seem to be any obvious avenues for repairing them. I should emphasize, however, that I do not think the counterfactual approach has nothing to offer in providing a *revisionary* metaphysics of ‘causation’. As I argued in chapters four and five, our ordinary concept of causation does not seem to be objective. It seems to be highly pragmatically-infected, and our causal judgments appear to be influenced by moral, intentional and epistemic factors. One might argue that such factors can have no place in an objective metaphysical picture of the world and seek to provide a revisionary account of causation. Hitchcock (forthcoming) distinguishes what he calls “token causal structure” (effectively token counterfactual structure) from token causation. The former is a fully objective feature of the world, consisting of the vast array of relations of counterfactual dependence, whereas the latter is subjective. Beebe (2004) also seeks an objective metaphysics of causation, consistent with the so-called “network model,”⁶¹ and acknowledges that such an account may have to be revisionary.

For the purposes of providing a descriptive conceptual analysis of our ordinary intuitions, however, I think that the Holding Fixed strategy is probably something of a red herring. Its primary advantage is that it enables counterfactual theories to deliver the intuitively correct verdict for cases of early and late preemption. But this move causes mayhem in other types of example, especially in cases of Switching and Self-Canceling Threats. If there were some *other* move that would enable us to capture preemption, we could avoid the Holding Fixed strategy altogether, and spare ourselves the huge headaches that it causes. Hall’s division of our concept of causation into two concepts is an exciting advance. It enables preemption to be treated as a local and intrinsic relation, without explicit reference to counterfactual dependence. Hall’s theory does have some counterexamples, but I feel they are of a much less serious kind than those facing the univocal counterfactual theorist. It seems as though Hall must have got *something* very right in his move towards pluralism. I have argued that there are even more senses of causation than the two that Hall posits. Hall’s theory does not have any obvious means for

⁶¹ The name “network model” due to Steward (1997).

generating indeterminate verdicts; it does not have the extra fluidity that a cluster theory has in this respect. The cluster theory is somewhat difficult to articulate; not in its logical form, but in specifying its content in any satisfying manner. It is all very well to say that C is a cause of iff *some* properties of the cluster are satisfied, but unless more detail is provided on *which ones*, it feels somewhat hollow. The ability of the cluster/prototype theory to explain prototype, however, counts hugely in its favor.

Three promising directions for further research that particularly come to mind. First, to attempt to fill out the content of a cluster-style analysis by more tightly constraining which combinations of properties suffice for a relation to be causal. This should proceed only so far, of course, since one of the basic commitments of the cluster approach is that there is a good deal of indeterminacy over which subsets of the cluster set/prototypical case are sufficient.

Second, the project of descriptive conceptual analysis would be greatly facilitated by the collection of systematic data on what our intuitions actually are. It seems extremely likely that experiments would reveal very interesting prototypical effects that would further illuminate our understanding of the concept.

Third, in response to the seemingly increasing rate of incursion of subjectivity and pragmatic factors into analyses of causation, it seems worth thinking more carefully about what *sort* of concept of causation we would like to have in order to carry out different philosophical and scientific tasks.

BIBLIOGRAPHY

- Alicke, M. 1992. Culpable Causation. *Journal of Personality and Social Psychology* 63: 368 – 378.
- Anscombe, G.E.M., 1971. Causality and Determination: An Inaugural Lecture. Cambridge: Cambridge University Press.
- Beebe, H. 2004. "Causing and Nothingness," in Collins, Hall and Paul (2004), pp. 291 – 308.
- Cartwright, N. 1979. "Causal Laws and Effective Strategies." *Noûs* 13: 419-437. Reprinted in How the Laws of Physics Lie (Oxford: Clarendon Press, 1983).
- Cartwright, N. 1983. How the Laws of Physics Lie, Oxford: Clarendon Press.
- Cartwright, N. 1989. Nature's Capacities and their Measurement, Oxford: Clarendon Press.
- Cartwright, N. 1999: The Dappled World, Oxford: Oxford University Press.
- Collins, J. 2000. "Preemptive Prevention," *Journal of Philosophy* 97: 223 – 234. Reprinted in Collins, Hall, and Paul (2004), pp. 107 – 117.
- Collins, J., N. Hall, and L. Paul, eds. 2004. *Causation and Counterfactuals*. Cambridge MA: MIT Press.
- Dowe, Phil 1992. "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory", *Philosophy of Science* 59: 195-216.
- Dowe, Phil 1996. "Backwards Causation and the Direction of Causal Processes", *Mind* 105: 227-248.

- Dowe, Phil 2000. Physical Causation. New York: Cambridge University Press.
- Eells, Ellery 1991. Probabilistic Causality. Cambridge: Cambridge University Press. Hall, N. 2004. "Two Concepts of Causation," in Collins, Hall and Paul (2004), pp. 225 – 276.
- Fair, David 1979. "Causation and the Flow of Energy", *Erkenntnis* 14: 219-50.
- Gasking, Douglas, 1955. "Causation and Recipes". *Mind*, 64, 479-87. Reprinted in Oakley and O'Neill 1996: 106-115.
- Gaut, B. 2000. "Art as a Cluster Concept". In Theories of Art. Carroll, N. ed.
- Godfrey-Smith, P. (forthcoming). "Causal Pluralism." Oxford Handbook of Causation. Hitchcock, Beebe, eds.
- Goodman, N. 1979. Fact, Fiction, and Forecast Cambridge MA: Harvard University Press.
- Hall, N. 2000. "Causation and the Price of Transitivity", *Journal of Philosophy* 97: 198-222.
- Hall, N. 2004. "Two Concepts of Causation", in Collins, Hall and Paul (2004).
- Halpern, J. and J. Pearl. 2001. "Causes and Explanations: A Structural-model Approach — Part I: Causes," *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence* (San Francisco: Morgan Kaufmann), pp. 194—202.
- Halpern, J. and J. Pearl. 2005. "Causes and Explanations: A Structural-model Approach — Part I: Causes" (expanded version), *British Journal for the Philosophy of Science*.
- Hart, H.L.A. and A.M. Honoré 1985. Causation in the Law (2nd Edition) Oxford: Clarendon Press.
- Healey, R. 1992. "Discussion: Causation, Robustness, and EPR" *Philosophy of Science* 59: 282-292.
- Healey, R. 1994. "Nonseparable Processes and Causal Explanation." *Studies in History and Philosophy of Science* 25 (3): 337-374.
- Hiddleston, E. 2005. "Causal Powers," *British Journal for the Philosophy of Science*.
- Hitchcock, C. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *Journal of Philosophy* 98: 273-299.
- Hitchcock, C. 2003. "Of Humean Bondage," *British Journal for the Philosophy of Science* 54: 1 – 25.

- Hitchcock, C. 2004b. "Do All and Only Causes Raise the Probabilities of Effects?" in Collins, Hall, and Paul (2004), pp. 403 - 417.
- Hitchcock, C. (forthcoming). "How to be a Pluralist about Causation."
- Hitchcock, C. MSa. "Token Causation."
- Hitchcock, C. MSb. "Prevention, Preemption, and the Principle of Sufficient Reason."
- Hume, David 1888 (first published 1739). A Treatise of Human Nature, ed. L.A. Selby-Bigge. Oxford: Clarendon Press.
- Hume, David 1902 (first published 1748). An Enquiry Concerning Human Understanding, ed. L.A. Selby-Bigge. Oxford: Clarendon Press.
- Knobe, J. MS. "Attribution and Normativity: A Problem in the Philosophy of Social Psychology."
- Lakoff, G. 1987. Women, Fire and Dangerous Things. Chicago: Chicago University Press.
- Leibniz 1716. Third Letter to Samuel Clarke, February 25, 1716. Reprinted in H. G. Alexander, ed. The Leibniz-Clarke Correspondence (Manchester: Manchester University Press, 1956), pp. 25 – 30.
- Lewis, David 1970. "How to Define Theoretical Terms", *Journal of Philosophy* 67: 427-446. Reprinted in Lewis 1983a: 78-95.
- Lewis, D. 1973. "Causation." *Journal of Philosophy* 70: 556 – 567. Reprinted in Lewis (1986b), pp. 159 - 172.
- Lewis, D. 1986a. "Postscripts to 'Causation'," in Lewis (1986b), pp. 159 - 213.
- Lewis, D. 1986b. *Philosophical Papers, Volume II*. Oxford: Oxford University Press.
- Lewis, D. 2000. "Causation as Influence." *Journal of Philosophy* 97: 182 - 197. Expanded version appears in Collins, Hall and Paul (2004), pp. 75 – 106.
- McDermott, M. 1995. "Redundant Causation," *British Journal for the Philosophy of Science* 46: 523 – 544.
- Mellor, D. H. 1995. The Facts of Causation. London: Routledge.
- Menzies, P. 2004b. "Causal Models, Token Causation, and Processes," *Philosophy of Science* 71 (Proceedings): S820 - S832.
- Mill, J. S. 1843. A System of Logic, London: Parker and Son.

- Pearl, J. 2000. Causality: Models, Reasoning, and Inference. Cambridge: Cambridge University Press.
- Salmon, W. C. 1984. Scientific Explanation and the Causal Structure of the World. Princeton: Princeton University Press.
- Salmon, Wesley 1994. "Causality Without Counterfactuals", *Philosophy of Science* 61: 297-312.
- Schaffer, J. 2000a. "Causation by Disconnection," *Philosophy of Science* 67: 285 – 300.
- Schaffer, J. 2000b. "Trumping Preemption," *Journal of Philosophy* 97: 165 – 181. Reprinted in Collins, Hall, and Paul (2004), pp. 59 – 73.
- Schaffer, Jonathan 2001. "Causes as Probability-Raisers of Processes", *Journal of Philosophy*, 98: 75-92.
- Sober, E. 1985. "Two Concepts of Cause" in P. D. Asquith and P. Kitcher (eds.) PSA 1984, Vol. II. East Lansing: Philosophy of Science Association, pp. 405 - 424.
- Steward, H. 1997. The Ontology of Mind, Oxford: Clarendon Press.
- Skyrms, B. 1980. Causal Necessity. New Haven: Yale University Press.
- Skyrms, B. 1984 "EPR: Lessons for Metaphysics" *Midwest Studies in Philosophy* 9: 245-255.
- Wittgenstein, L. 1953. Philosophical Investigations.
- Woodward, J. 2003. Making Things Happen: A Theory of Causal Explanation, Oxford: Oxford University Press.
- Yablo, S. 2002. "De Facto Dependence," *Journal of Philosophy* 99: 130 – 148.
- Yablo, S. 2004. "Advertisement for a Sketch of an Outline of a Prototheory of Causation," in Collins, Hall and Paul (2004), pp. 119 – 137.