Causation and Its Basis in Fundamental Physics

Douglas Kutach

Final Version

{ CONTENTS }

Preface xi

1. Empirical Analysis and the Metaphysics of Causation 1

1.1 Empirical Analysis 2

1.1.1 The Distinctive Features of Empirical Analysis 9

1.2 Empirical Analysis of the Metaphysics of Causation 13

1.2.1 Effective Strategies 14

1.3 Empirical Analysis of the Non-metaphysical Aspects of Causation 17

1.4 Causation as Conceptually Tripartite 20

1.5 A Sketch of the Metaphysics of Causation 22

1.6 Fundamental and Derivative 24

1.6.1 The Kinetic Energy Example 25

1.6.2 Some Constitutive Principles of Fundamentality 30

1.7 Abstreduction 34

1.8 STRICT Standards and RELAXED Standards 37

1.9 Limitations on the Aspirations of Empirical Analysis 44

1.10 Comparison of Empirical and Orthodox Analysis 45

1.11 Summary 48

PART I **The Bottom Conceptual Layer of Causation**

2. Fundamental Causation 53

2.1 Preliminaries 57

2.1.1 Events 58

2.1.2 Laws 64

2.2 Terminance 67

2.2.1 Causal Contribution 75

2.2.2 Trivial Terminance 76

2.3 The Space-time Arena 79

2.4 Classical Gravitation 82

2.4.1 Galilean Space-time 82

2.4.2 Terminants in Classical Gravitation 83

2.4.3 Overdetermination in Classical Gravitation 85

2.4.4 Instantaneous Causation 87

2.5 Relativistic Electromagnetism 89

2.5.1 Minkowski Space-time 89

2.5.2 Terminants in Relativistic Electromagnetism 91

2.5.3 Classical Unified Field Theory 92

2.6 Content Independence 94
2.7 Continuity and Shielding 98
2.8 Transitivity 100
2.9 Determinism 101
2.10 Stochastic Indeterminism 104
2.10.1 Stochastic Lattices 104
2.10.2 A Toy Theory of Particle Decay 105
2.11 Non-stochastic Indeterminism 107
2.11.1 Newtonian Indeterminism 107
2.11.2 Contribution Extended 109
2.12 General Relativity 109
2.12.1 Spatio-temporal Indeterminism 110
2.12.2 Closed Time-like Curves 111
2.13 Quantum Mechanics 112
2.13.1 The Quantum Arena and its Contents 112
2.13.2 Bohmian Mechanics 113
2.13.3 Spontaneous Collapse Interpretations 114
2.13.4 Other Interpretations of Quantum Mechanics 116
2.14 Summary 116

PART II **The Middle Conceptual Layer of Causation**

3. **Counterfactuals and Difference-making** 121
3.1 General Causation 121
3.2 Counterfactuals 123
3.3 Goodman's Account of Counterfactuals 124
3.4 The Nomic Conditional 128
3.5 Comparison to Ordinary Language Conditionals 134
3.6 Prob-dependence 135
3.7 Contrastive Events 136
3.8 Summary 138

4. **Derivative Causation** 140
4.1 Influence 140
4.2 Prob-influence 142
4.3 General Causation 144
4.4 Temporally Extended Events 150
4.5 Idiomatic Differences between Promotion and Causation 152
4.6 Aspect Promotion 154
4.7 Promotion by Omission 155
4.8 Contrastivity 155
4.9 Transitivity 157
4.10 Continuity 160
4.11 Shielding 161
4.12 Partial Influence 162
4.13 Summary 167

5. The Empirical Content of Promotion 168

5.1 The Promotion Experiment 169

5.2 Insensitivity Considerations 171

5.3 Thermodynamics and Statistical Mechanics 175

5.4 The Asymmetry of Bizarre Coincidences 184

5.5 The Analogy to Thermal and Mechanical Energy 186

5.6 Broad and Narrow Promotion 187

5.7 Inferences from Empirical Data to Promotion 193

5.7.1 Simpson's Paradox 194

5.8 Why There Are Effective Strategies 197

5.9 Mechanistic Theories of Causation 199

6. Backtracking Influence 203

6.1 The Direction of Influence 205

6.2 Proof of Causal Directness 207

6.3 A Search for Empirical Phenomena 214

6.4 'Past-directed then Future-directed' Influence 219

7. Causal Asymmetry 225

7.1 The Empirical Content of the Causal Asymmetry 226

7.2 Causation and Advancement 230

7.3 An Explanation of the Advancement Asymmetry 232

7.3.1 Prob-influence through Backtracking 234

7.3.2 Directly Past-directed Prob-influence 236

7.3.3 Summary 238

7.4 Pseudo-backtracking Prob-influence 239

7.5 The Entropy Asymmetry and Causal Directionality 244

7.6 Recent Alternative Explanations of Causal Asymmetry 249

7.6.1 The Albert-Kutach-Loewer Approach 250

7.6.2 The Price-Weslake Approach 252

7.6.3 The Fork Asymmetry Approach 254

7.7 Fundamental Influence Asymmetry 257

7.7.1 By Fiat 258

7.7.2 By Happenstance 260

7.8 Summary 262

PART III The Top Conceptual Layer of Causation

8. Culpable Causation 265

8.1 The Empirical Insignificance of Culpability 270

8.1.1 Part I: Singular Causation 271

8.1.2 Part II: General Causation 272

8.2 Culpability as a Heuristic for Learning about Promotion 275

8.3 Culpability as an Explanatory Device 277

8.4 Culpability as a Proxy for Terminance and Promotion 277

8.5 Commentary 281

9. The Psychology of Culpable Causation 282

9.1 The Toy Theory of Culpable Causation 284

9.2 Culpability$_1$ 287

9.2.1 Salience 287

9.2.2 Irreflexivity 289

9.2.3 Asymmetry 289

9.2.4 Significant Promotion 290

9.3 Shortcomings of Culpability$_1$ 291

9.3.1 Precise Character of the Effect 291

9.3.2 Overlapping Causation 291

9.3.3 Probability-Lowering Causes 292

9.4 Culpability$_2$ 292

9.5 Shortcomings of Culpability$_2$ 297

9.5.1 Saved Fizzles 298

9.5.2 Early Cutting Preëmption 298

9.5.3 Late Cutting Preëmption 298

9.6 Culpability$_3$ 299

9.7 Culpability$_4$ 300

9.8 Summary 305

10. Causation in a Physical World 307

10.1 Summary 307

10.2 Future Directions 312

References 321

Index 329

{ 1 }

Empirical Analysis and the Metaphysics of Causation

It is common knowledge that no one understands all the causes and effects that occur in nature, but it may surprise the uninitiated that the idea of causation itself—what causation amounts to—is thoroughly contested among experts without anything remotely approaching a consensus as to the likely form of a satisfactory account. The lack of an accepted theory of the connection between cause and effect coexists with a general agreement that some sort of causation is a critical component of reality. Causation is undoubtedly important in science, but even subjects like ethics, politics, and theology are significantly constrained by the need to make their proclamations compatible with what we know about paradigmatically causal interactions among ordinary physical objects. Causation's central role in linking various components of our overall worldview is evident in the wide range of theories that rely on the coherence of some notion of causation to account for perception, names, time, knowledge, and more.

Satisfying the full range of desiderata for an account of the metaphysics of causation has proven a stiff challenge, illustrated by an expansive technical literature. As I see it, the traditional approaches that dominate philosophical discussion have already succeeded in identifying virtually all the important elements needed for a comprehensive understanding of causation. However, the productive components have not yet been assembled into a convincing systematic metaphysics of causation because the traditional conception of what a metaphysics of causation is supposed to do—provide an informative and principled and consistent regimentation of important truths about causation—virtually ensures failure.

Fortunately, there exists an alternative conception of the task a metaphysical account of causation ought to accomplish: empirical analysis. A successful empirical analysis would vindicate enough of our use of causal concepts in science and philosophy and ordinary life in order for us to claim success in understanding the metaphysics of causation. Empirical analysis redraws the boundaries of the conceptual geography in a way that makes an adequate metaphysics of causation much easier to construct, in effect lowering the bar for success.

For illustration, one can consider the "problem of preëmption" that is believed to plague some prominent theories of causation. Preëmption is when a potential cause is on the way toward producing an effect but is somehow forestalled, like a lit fuse that is severed before it can ignite its rocket. Traditionally, the metaphysics of causation has been understood as needing to provide rules that identify paradigmatic preëmpted would-be causes as not genuine causes. In an empirical analysis of causation, however, it turns out that preëmption does not need to be understood as part of metaphysics, which permits it to be addressed successfully according to more lenient standards. An empirical analysis of the metaphysics of causation can safely count paradigmatic preëmpted would-be causes as bona fide causes.

My task in this book is to explicate this empirical approach and to implement its methodology in constructing a comprehensive metaphysics of causation. Along the way, I will dedicate a great deal of attention to how fundamental physics can serve as a basis for an adequate metaphysics of causation. However, readers are here warned that it is far beyond the scope of this book for me to argue that causation can be reduced to fundamental physics or that causation in the special sciences is merely a bunch of physics. I have much to say on this topic, but to address this important issue properly, I first need to set out my account of causation and the methodology behind it. That alone is a substantial enough task for a single volume.

This introductory chapter consists of two main components: an explanation of the non-standard methodology I will employ and a sketch of the overall structure of causation according to my account. In the first half of this chapter, I will provide a preliminary exposition of empirical analysis and an explanation of how it differs from orthodox conceptual analysis. Then I will illustrate how empirical analysis applies specifically to the metaphysics of causation and to the non-metaphysical aspects of causation. In the second half, I will describe how causation can be distinguished into three stacked conceptual layers. In order to clarify the three layers, I will need to unpack the two distinctions that mark the boundaries between them. One distinction is between fundamental reality and derivative reality, and the other is between STRICT and RELAXED standards of theoretical adequacy. Using the new terminology, I will summarize my account of causation and outline how the remaining chapters will address it in more detail.

1.1 Empirical Analysis

I have chosen 'empirical analysis' as the label for the methodology I will be employing throughout my investigation of causation as a figurative tip of my hat toward Phil Dowe's (2000, Ch. 1) discussion of conceptual analysis and its application to causation. I will soon explain what I mean by 'empirical analysis', but a brief warning is likely warranted for readers familiar with Dowe's work. The version of empirical analysis I will adopt is consistent with what Dowe says about empirical analysis, but I impose further conditions on what constitutes a proper empirical analysis that are substantial enough to make it incorrect to equate my philosophical project with Dowe's. It will turn out, for example, that my metaphysics of causation does not compete with theories where causation is understood in terms of the transfer of physical quantities or the paths of particles bearing conserved quantities. Although such theories are commonly understood as empirical approaches to causation, no existing examples from this tradition count as empirical analyses of causation under my narrower conception of the methodology. I can even go so far as to say I am unaware of any published example of what I would categorize as an empirical analysis.

Because others have not endorsed my construal of empirical analysis, it would be inappropriate for me to contrast other prominent accounts with mine in an effort to assess their relative merits. As I emphasized in the preface, if I were to criticize some account for being inadequate according to the standards of empirical analysis, it would be all too easy for the author to mount an effective defense by simply denying that the targeted account was ever intended to be an empirical analysis (of the form I have defined). Yet, if I were to set aside the method of empirical analysis, I would not have much to say that would still be relevant to the metaphysical project being conducted in this volume. In §5.9, I will clarify why currently existing transference and causal process theories like Dowe's do not count as empirical analyses according the precisification I am invoking.

Let us now attend to a positive characterization of empirical analysis. Uncontroversially, it proves useful in science to employ specialized terminology that is honed to improve precision, simplicity, and generality. To conduct the empirical analysis of some topic *X*, overly broadly speaking, is to identify scientifically improved concepts of *X*. Empirical analysis is a form of conceptual analysis in the broad sense that it provides a link between our ordinary conception of *X* and things in the world, but it is a non-standard form of conceptual analysis by forging the linkage in a manner especially responsive to scientific theorizing and experimental results. Empirical analysis involves not merely setting aside disagreements between theoretically refined terms and the platitudes that characterize *X*, which is common in contemporary versions of conceptual analysis (Jackson 1998), but also adapting the refined terminology to improve explanations of *experiments* that conceptually encapsulate the empirical phenomena that make our concept of *X* worth having.

In my experience, the distinctive features of empirical analysis are surprisingly difficult for experienced philosophers to grasp. Readers are thus cautioned not to be too hasty in concluding that they fully understand what constitutes a proper empirical analysis because some important clarifications cannot be adequately stated until §1.9 after I have introduced some new terminology.

One can acquire a decent preliminary grasp of empirical analysis by reviewing how exemplary sciences engineer their conceptual schemes. For example, food scientists are interested in answering questions about why some foods are healthier than others. A scientific investigation of food provides explanations for the following kinds of empirical phenomena. People who eat a mixture of fruits, vegetables, and nuts are healthier, ceteris paribus, than people who eat sand. The design of concepts for food science ought to be honed toward maximizing the quality of such explanations by having a regimentation of our folk food concept that strikes an adequate balance among various dimensions of explanatory quality such as simplicity, capturing as many empirical phenomena as possible within the scope of the explanation, and fitting properly with related subjects like agronomy, physiology, and chemistry. As it turns out, food science does have an improved concept of food, which we commonly refer to as 'nutrient'. 'Nutrient' serves as an excellent substitute for 'food' when studying the health effects of various ingested substances in part because it is only loosely tethered to our ordinary food concept. Crucially, we do not want to reject a theory of nutrition because it identifies iron crowbars, dust mites, and oxygen as nutrients whereas folk opinion adamantly rejects these as foods. (It ought to go without saying that it is altogether irrelevant that by historical happenstance English has two etymologically distinct words—'food' and 'nutrient'—for the folk and scientific concepts, respectively. There are abundant examples of the same word being used in an informal sense and as a technical term.)

To optimize our native food concept in the service of food science is to make it more precise in a way that achieves an optimal or at least acceptable level of quality according to principles of good conceptual design and according to the needs of food science. A precisification of 'food' or 'nutrient' can be thought of as a stipulation of a maximally precise class of all the possible entities that count as nutrients. I will call each class an 'intension' of the concept.

I will now clarify a few issues relevant to conceptual design and its role in empirical analysis. I will not attempt to provide a complete list of principles for engineering concepts nor a specification of their relative importance, but I will instead assume current scientific practice serves well enough for guidance.

First, philosophers have debated whether we should think of a conceptual analysis of *X* as trying to make an a priori claim about what *X* must be, given how our concept of *X* works or whether we should allow the analysis to incorporate some a posteriori component, for example (Block and Stalnaker 1999, Chalmers and Jackson 2001). From the perspective of empirical analysis, this way of framing the status of conceptual analysis does not elucidate the key issue. The purpose of empirical analysis is not to evaluate our mutually shared folk concept *X* in detail from the armchair, but to take what data science provides and to organize that data from the armchair to arrive at superior surrogates for *X*. The primary task is to balance the generality and specificity of the sought-after scientifically honed concepts. For illustration, consider one of Phil Dowe's glosses on an empirical analysis of causation: that it is intended to "discover what causation is in the objective world" (2000, p. 1). Such a project does not require settling questions about causation in all conceivable worlds, like worlds where magical spells are operative or where time does not exist.

Dowe's approach is rightly criticized by Collins, Hall, and Paul (2004, p. 14) for offering an improperly narrow treatment of the connection between fundamental laws and causal facts. They claim it would be better to "specify the way in which the fundamental laws fix the causal facts in terms that *abstract away* from the gory details of those laws—thereby to produce an account that has a hope of proving to be not merely true, but necessarily so."

I think the correct way to resolve this dispute is to recognize that there are two competing virtues, neither of which should dominate the other. On the one hand, we should prefer our concepts to be insensitive to the details of any empirical phenomenon we have not yet figured out. The nineteenth century conception of energy, for example, is largely insensitive to the details of microscopic interactions. Our later discovery that there is a strong and a weak nuclear interaction did not force a revision of that concept of energy or its central applications—for example, to the feasibility of constructing perpetual motion machines—because the concept of energy was already sufficiently insulated from such details. That counts as a successful application of conceptual engineering.

On the other hand, we should care little about how our concept applies to highly unrealistic possibilities and not at all about whether it applies to absolutely every possibility. It is patently silly, for example, to require a scientific theory of energy to accord with pre-theoretical intuitions we may have concerning the energetic consequences of magic spells. An important rule of good conceptual design is to avoid optimizing concepts to better handle epistemically remote possibilities when it proves costly to the explanation of more realistic possibilities. How we conceive of magic can bear on empirical analysis by helping to clarify how our concepts generalize, but as the imagined possibilities become ever more outlandish, there is less need to fiddle with our concepts in order to accommodate people's gut intuitions. Empirical analysis should not be adapted merely to what we currently believe to be true about the actual world, but neither should it be required to accord with everything we naturally want to say about every conceivable possibility.

Second, a concept can be virtuous by being appropriately insensitive to details that are unimportant in application. For example, whether S should count as a nutrient should be insensitive to whether S is nutritive in its present condition or only after further chemical changes that will occur during cooking or digestion. An important special case of this principle is that it is virtuous for one's postulated conceptual relations to harmonize with each other and exhibit *graceful degradation* when the applicability of one concept breaks down. For example, an empirical analysis of food ought to be compatible with the observation that there are borderline cases of nutrients and cases where a substance is slightly nutritional in one respect yet slightly poisonous in another. An empirical analysis would be deficient if it required a definite binary fact of the matter about whether S is a nutrient even though classical logic requires S to be either a nutrient or not.

The importance of engineering the graceful degradation of concepts can be illustrated in terms of the conserved quantity (CQ) account of causation (Dowe 1992a, 1992b, 2000), which postulates that causal interaction requires the transfer of some conserved quantity. Suppose it turns out that the actual fundamental laws ensure that quantities like energy, momentum, and angular momentum are always conserved so that the CQ theory is applicable to the actual world. We can ask about what would have been true about causation if the laws of nature were very slightly adjusted so that conserved quantities became very nearly but not perfectly conserved in a way that preserved all the actual world's macroscopic regularities, the regularities that give us a good reason to believe in causation. It is a consequence of the CQ theory that there would be no causation in such a world. That result by itself may be acceptable, but we have a right to expect the CQ theory to provide some account of why the complete lack of causation coexists with the vast evidence we would have for causation. According to my version of empirical analysis, the provision of such a story needs to be part of the CQ theory's explanation of why the conservation of quantities matters. It is unacceptable for the CQ theorist to balk that because conservation holds in the actual world, consideration of alternative worlds where it does not hold is irrelevant to the study of causation in the actual world. It is relevant to the analysis because if causation requires truly conserved quantities, then it becomes mysterious how we could ever become aware of causal connections, for we are likely not in a position to tell whether nature's quantities are perfectly conserved or just very nearly conserved. The CQ theorist needs to provide some explanation of how we could have epistemological access to causal relations. One schema for a proper explanation would involve demonstrating that our evidence for the existence of causal relations depends on how closely a universe obeys a conservation law. If it could be shown that putative evidence for causation becomes progressively weaker as violations of conservation laws accumulate in number or magnitude, then the CQ theorist could argue that even though worlds with only close approximations to conserved quantities have no genuine causation, we are reasonable to interpret them as having causation when they obey the appropriate conservation laws so far as we can tell. I am not contending that this particular explanation is satisfactory for CQ accounts, only that some story needs to be given about how breakdowns in the applicability of the concepts used in the empirical analysis relate to breakdowns in the applicability of the target concept.

Third, empirical analysis appears to presuppose a distinction between that which is empirical and that which is not. If this distinction is too narrowly construed, problems arise. In a vast array of examples, things we naïvely take to be unproblematically observable turn out to be characterizable only in theoretically loaded language. Also, we can often shift seamlessly between what counts as observed and what counts as inferred. When I claim to see a sheep on the hill, am I seeing a sheep or am I seeing half of a sheepish surface and inferring the rest of the sheep, or am I seeing a colored patch and inferring from that? There appears to be quite a bit of flexibility in how we can answer that question. In order to bracket concerns about how principled the concept 'empirical' is, I intend 'empirical analysis' to be understood not to presuppose a determinate fact of the matter about which items are genuinely empirically accessible. Instead, we should require a successful empirical analysis to make claims that are suitably insensitive to any indeterminacy concerning the empirical. This bracketing will not answer any probing epistemological questions, nor will it ensure the existence of a sufficiently principled empirical basis for adjudicating among competing empirical analyses. However, such deferral is common throughout science. So, to any worries that my empirical analysis of causation requires an unreasonably clear distinction between the empirical and non-empirical, my response is simply that my implementation of the distinction is no different from what is employed throughout science.

Fourth, a principle that is crucially not a part of empirical analysis is a preference for the intension of the analyzed concept to coincide with folk opinions about paradigm cases. It is literally of zero importance for an empirical analysis that paradigm instances of food count as nutrients. If the candidate intension for 'nutrient' happens to count bread as a non-nutrient, that by itself does not count as a deficiency, no matter how strong our pre-theoretical commitment to the proposition that bread is a foodstuff, and no matter how large a fraction of the general public supports the proposition that it is obvious that bread is food, and no matter how many professors one can summon to assert expert testimony that, a priori, bread is food. There does need to be enough of a semantic connection between paradigm cases of food in aggregate and the intension of 'nutrient' so that one cannot pass off an unrelated concept as a theoretical refinement of 'food'.[1] Nevertheless, the connection between the consequences of the theoretical refinement and the original platitudes can permit abundant disagreements without at all detracting from the quality of the conceptual regimentation.

In order to get a better grasp of this contrarian principle, it may help to consider the concept of rotation. Anyone interested in understanding the rotation of material objects is well served by group theory, the branch of mathematics designed to characterize symmetries. There is a group, for example, that represents the relations between all the possible rotations an object can undergo in a two-dimensional plane around a single point. The members of this group can be represented by real numbers. The number θ corresponds to a counter-clockwise rotation by θ radians. Negative numbers correspond to clockwise rotations, and the zero rotation corresponds to no rotation at all. If we were to apply the principle that a conceptual analysis of rotation should make explicitly true those propositions that are analytically true of our folk concept, then we would need to judge that the group-theoretical conception of rotation is in some measure deficient because it counts a rotation of zero as a bona fide rotation. What could be a more paradigmatic non-rotation than something that rotates a zero amount? The reason zero rotations are included in

[1] One could argue that CQ theories of causation are inadequate for this reason, for contemporary versions do not adequately explain why the transmission of conserved quantities is relevant to the causal principles successfully used in the special sciences and in everyday life.

the group-theoretic concept is that it greatly simplifies the definitions and theorems concerning relations among different kinds of rotation. For example, we would like to be able to say that the composition of any two rotations is itself always a rotation, but we cannot state that claim with optimal simplicity if zero rotations are forbidden because a rotation by θ and then by $-\theta$ amounts to a net non-rotation. Mathematicians understand the zero rotation as a trivial rotation rather than something that is not a rotation at all. This respectable attitude is strikingly at odds with the kind of conceptual analysis typically assumed in modern discussions of the metaphysics of causation. It is often taken for granted that events do not cause themselves and that a satisfactory metaphysics of causation needs to accord with this truth by not making it explicitly true that every event causes itself. According to the standards appropriate to empirical analysis, however, it is perfectly acceptable for a metaphysics of causation to ensure that every event causes itself. One can dismiss the importance of this counterintuitive result merely by recognizing that self-causation is a trivial form of causation.

For future reference, I will continue to use the word 'explicitly' in statements of the form, "Theory T (or model M) makes P explicitly true," to communicate that T (or M) suffices for P in the most straightforward interpretation of its claims, setting aside adjustments for language pragmatics. I have just provided two examples of what I mean by 'explicitly' in this context. The mathematician's regimentation of 'rotation' in terms of group theory makes explicitly true that an object at rest is undergoing rotation. One regimentation of 'cause' I will advocate makes explicitly true that every event causes itself. In an empirical analysis, it is no knock on a theory that it makes explicitly true claims we know are false because such discrepancies can be harmlessly explained away in terms of language pragmatics.

Although empirical analysis is largely an activity of regimenting concepts that can be conducted from one's philosophical armchair, the ultimate goal is not the investigation of language or thought but finding the best scientific theory one can. In an empirical analysis of food, the data one seeks to systematize are primarily all the statistical correlations between an animal's biological condition together with what it ingests and its later health condition, but other kinds of data are also relevant. What ought to concern us is learning about robust regularities in these data. The system of concepts provided by an empirical analysis plays a housekeeping role, keeping the conceptual system functioning as efficiently as feasible. Although empirical analysis is subservient to science, that does not trivialize the activity of finding an adequate empirical analysis. For one thing, trying to optimize one's conceptual scheme can play an instrumental role in science. It can raise possibilities that would not otherwise be entertained and can identify some issues as pseudo-problems. My explanation of causal asymmetry in chapter 7 provides an instructive example. For another thing, as Wilfrid Sellars (1962) put it, philosophy aims to find out "how things in the broadest possible sense of the term hang together in the broadest possible sense of the term." Understanding how things hang together is largely a project of conceptual engineering.

1.1.1 THE DISTINCTIVE FEATURES OF EMPIRICAL ANALYSIS

In order to highlight the novel features of empirical analysis, I will now contrast it with what I will call 'orthodox conceptual analysis', or just 'orthodox analysis' for short. Unfortunately, owing to space limitations and the inherent difficulty of criticizing the murky methodology of orthodox analysis in a manner sufficiently resistant to misinterpretation, I can only comment briefly. I have engaged this topic previously in (Kutach 2010), and I have made additional commentary publicly available for interested readers to follow up on this topic in more detail.

A conceptual analysis of *X*, as I will understand it here, is a systematization of the platitudes that constitute our implicit concept of *X*. To conduct a conceptual analysis, one starts with some initial data in the form of uncontroversial truths about the concept, including exemplars of the concept as well as a priori links to other concepts. For example, a conceptual analysis of food would begin with propositions that an orange is food, a hoagie is food, etc., as well as with broader principles that food is the kind of thing people typically like to eat, the kind of thing that relieves hunger, a kind of material substance, a kind that is species-relative, etc. One then attempts to formulate a reasonably small set of principles that (perhaps with some innocuous auxiliary truths) implies a set of claims that comes close enough to matching the initial platitudes. This set constitutes the completed conceptual analysis. It is understood that such a conceptual regimentation can be acceptable and even exemplary even when it rejects the truth of some of the initial platitudes. Indeed there are a wide variety of stances one can take on which kinds of discrepancies between the consequences of the completed conceptual analysis and the initial platitudes are permissible for the conceptual analysis to count as successful.

Some philosophers steeped in the naturalist tradition may think that conceptual analysis has long ago been abandoned, and they would be correct insofar as we understand conceptual analysis narrowly in its old-fashioned forms like Curt John Ducasse's(1926) attempt to define the causal relation. But the more liberal versions described by Collins, Hall, and Paul (2004) are currently in widespread use and have been prominently defended (Jackson 1998).

What makes a conceptual analysis *orthodox*, as I understand it, is the lack of any further systematic method (beyond custom, personal preference, appeasing journal referees, and the like) for adjudicating which discrepancies are acceptable for a satisfactory analysis and assessing the relative merits of competing analyses that differ in how well their accounts match the target platitudes. The implicit conditions of adequacy for orthodox analysis vary quite a bit among those who practice it, but a recurring feature of debates over whose analysis is adequate is the lack of any analysis that perfectly fits the initial platitudes and a proliferation of seeming stalemates among partially successful theories.

There are numerous examples in the philosophical literature on causation where two competing theorists appear to agree on all the relevant facts but disagree on how to incorporate them into an orthodox analysis. One good example is the transitivity of causation. In a scenario to be discussed later, Jane removes her food from a bear box and thus causes her food to be left in the open where bears can get it. Her food being left out causes Jill to recognize the danger and put the food back in the bear box. But Jane's removing the food from the bear box intuitively does not cause the food to be in the bear box later. Some investigators respond by denying the transitivity of causation. Others maintain the transitivity of causation and insist that Jane really does cause the food to be in the bear box but that we do not ordinarily identify such cases as causation because of pragmatic factors. Both sides can agree on all the relevant facts—which interactions occur, what the relevant laws are, which events affect the probabilities of other events—but still disagree about whether Jane's removal of the food was a cause of the food being in the box later. Chris Hitchcock(2003) has conveniently provided a discussion of many such quandaries for orthodox analyses of causation. One upshot is that orthodox conceptual analysis does not prescribe any discernible guidance for adjudicating such disputes and for assessing the relative importance of each initial platitude.

Empirical analysis is crucially different by incorporating specific additional methodology to *guide* movement away from the initial platitudes. Empirical analysis takes the platitudes concerning *X* as a starting point for identifying empirical phenomena, especially by formulating explicit experiments whose results clarify why *X* has some conceptual utility. Then, one's goal is to seek a scientific explanation for the results of such experiments, honing the concepts used in the explanation as much as needed to improve its overall quality, including how it comports with other background theories we accept. Whatever concepts result from this optimization constitute the completed empirical analysis of *X*.

An empirical analysis often results in some of the original platitudes being discarded as irrelevant to the analysis, and the final regimented concept is not to be evaluated in terms of what fraction of the platitudes it makes explicitly true. While orthodox analyses continue to be tethered to some extent to the initial platitudes by always being evaluated in the end in terms of the magnitude and severity of discrepancies with the initial platitudes, the method of empirical analysis encourages us to abandon the platitudes whenever making them explicitly true would result in a suboptimal conceptual scheme.

To encapsulate, empirical analysis may be given the following formal definition. The ***empirical analysis*** of *X* is the engineering of a conceptual framework optimized in the service of the scientific explanation of whatever empirical phenomena motivate our possession of a concept of *X*, especially insofar as they are characterized in terms of experiments.[2]

[2] In claiming that an empirical analysis of *X* addresses empirical phenomena motivating "our possession of a concept of *X*" I am referring to whatever concept (or perhaps concepts) we possess before we improve our conceptual scheme scientifically. Our rough and ready folk conception of *X* can be understood as having a very low threshold for being motivated or useful or worth

It ought to go without saying that scientists have been engaging in the activity of empirical analysis for centuries. In fact, the only argument I offer for empirical analysis being an acceptable form of conceptual analysis is that it has been conducted by scientists in countless instances, and its successful applications have greatly contributed to our understanding of reality. In this sense, there is nothing new about my approach to causation.

It also ought to go without saying that philosophers have long recognized that traditional forms of conceptual analysis include a role for scientific inquiry. Sometimes this idea is expressed as the claim that meanings are not entirely "in the head" (Putnam 1975), an observation illustrated by natural kind terms like 'water'. Not everything that behaves like paradigmatic water is water. Only substances of the same chemical kind as the chemicals that predominate among most paradigmatic instances of water in our local environment count as water. A non-H_2O chemical on the other side of the universe that behaves superficially like water is not water. That we can recognize this feature of our 'water' concept from our philosophical armchairs demonstrates that sometimes the intension of a concept depends on the external environment. Again, this is all well known; the mere incorporation of scientific discoveries into a conceptual analysis is also not a novel feature of empirical analysis.

What is new about empirical analysis—to philosophers, as far as I can tell—is the starring role it casts for explicitly characterized experiments. Later in this volume, I will attempt to construct three experiments in an effort to characterize the empirical phenomena giving us some reason to believe in causation: the promotion experiment, the backtracking experiment, and the asymmetry experiment. Two more examples of how empirical analysis relies on characterizing experiments can be found in (Kutach Forthcoming).

To grasp the crucial role of experiments in an empirical analysis, we can again consider the investigation of food. In an empirical analysis of food, one should attempt to describe a general experiment that captures the empirical phenomena that make 'food' a concept worth having. The following experimental schema, I think, serves reasonably well for a simplistic illustration. One chooses some type of creature, C, some type of edible material M, and some type of environment E for the creature to inhabit during the study. Then, one conducts an experimental run by having a chosen creature ingest a substance and measuring its health outcomes after its stay in the environment. After zillions of such experimental runs for a wide range of creatures, materials, and environments, one will have collected data that can be summarized as a function from these three variables to a set of health outcomes. The results of such experiments presumably verify that

possessing. One might say that any concept in regular use very likely has some value and is thus worth possessing, for otherwise it probably would have been abandoned. It is certainly possible that scientific investigation or empirical analysis will justify abandoning the folk concept or replacing it with alternatives. Contrary to Peter Godfrey-Smith's (2012) suggestion, empirical analysis does not require that our initial concept X retain its utility after we have engineered superior replacements for them.

humans eating sand and nothing else for a month results in bad health outcomes while eating vegetables, nuts, and fruits mostly results in good health outcomes. The reason our food concept is worth having is that there are robust regularities where certain kinds of ingested materials significantly improve health outcomes relative to other materials. We use 'food' primarily to track these nutrients, and completing the empirical analysis of food requires us to hone 'food' more precisely (using the label 'nutrient', if desired, to avoid the usual connotations of 'food') so that it fits better with everything else we know about physiology, chemistry, disease, and related subjects.

Because there are secondary factors bearing on our use of 'food' like its role in social encounters and its aesthetic qualities, there are mismatches between our judgments about which substances are food and what our science identifies as nutritional. The secondary factors bear on empirical analysis only through a much different set of empirical phenomena: primarily, people's reports about what they consider to be food. These empirical phenomena can be encapsulated in terms of an experiment, a decent first approximation of which would simply involve presenting a sample of material M to a human subject C in environment E and ask, "Is this food?" You could augment the experiment by also testing whether the person eats the substance or serves it to others at dinner, but the basic idea is to test not the bodily effects of the ingested substance but instead how people think of it, talk about it, and use it socially. The data collected from such experiments would constitute empirical phenomena concerning how we conceive of food, and this concept can be made more precise in order to explain why we have the intuition that microbes and humans and aspirin tablets are not food.

The first kind of empirical analysis is typically of much greater philosophical importance because it bears directly on the character of reality generally rather than focusing on how we conceive of it. My main reason for discussing the second kind of empirical analysis is to avoid alarming readers who insist there must be some place in our conceptual scheme for widely shared and strongly held intuitions about important concepts like causation. The intuitions that are properly ignored in an empirical analysis of the metaphysics of causation always have a proper home in this second more psychologically oriented empirical analysis. (Readers who are pressed on time and are untroubled by the fact that my metaphysics of causation does not address several folk intuitions that are traditionally construed as data for a metaphysical investigation of causation should be able to skip chapters 8 and 9.)

An exploration of food using the methodology of empirical analysis thus leads naturally to two somewhat separate investigations: identifying experiments that capture the nutritional aspects of food and identifying experiments that capture the social and psychological aspects of food. This sort of bifurcation happens quite generally when the method of empirical analysis is applied. In effect, empirical analysis attempts to provide with two analyses what an orthodox analysis attempts to accomplish with a single analysis. Typically, an orthodox conceptual analysis of some *X* starts with a set of platitudes concerning *X*, some of which are partly constitutive of the meaning of *X* and some of which are known from empirical investigation. Conducting an orthodox analysis of *X* consists in systematizing all these platitudes concerning *X* together as a single group by identifying a cluster of principles that describe what *X* is in terms of other concepts. In conducting an empirical analysis, by contrast, it usually works out that the original platitudes are best segregated into two groups, those bearing on *X* insofar as it is something "out there in reality" and those bearing on how we think about *X* in ways that go beyond the empirical phenomena in the first group. Then, one conducts two distinct regimentation projects.

Application of the methodology of empirical analysis to causation results in a natural bifurcation into a pair of empirical analyses. The first empirical analysis is more focused on causation insofar as it is something "out there in reality." I will refer to this investigation as the *empirical analysis of the metaphysics of causation*, or sometimes just the *metaphysics of causation*. The second empirical analysis focuses on how we think about causation in ways that go beyond the empirical phenomena addressed by the metaphysics of causation. I will refer to this investigation as the *empirical analysis of the non-metaphysical aspects of causation*. This second empirical analysis will subsume the psychology of causation as well as the subset of epistemology that encompasses the explanatory role of causes and causal modeling. Because this volume is primarily concerned with the metaphysics of causation, this second empirical analysis will receive far less attention from me than the first. My goal will be limited to sketching very briefly a few of its components just to assure readers that the topics it concerns are not being denigrated by my metaphysics of causation or entirely ignored but are merely being reorganized in a way that categorizes them as part of the special sciences themselves rather than as part of metaphysics. This conceptual division is not made for the sake of presentation but for the purpose of assigning to each empirical analysis the standards of theoretical adequacy that are appropriate to it.

Before further clarifying how I think of the proper standards of theoretical adequacy, I will provide some additional detail about the two empirical analyses of causation.

1.2 Empirical Analysis of the Metaphysics of Causation

The purpose of this section is to sketch how the general methodology of empirical analysis will yield my account of the *metaphysics of causation*. Later in §1.10, I will further clarify the character of this empirical analysis by defending an important restriction on the proper scope of metaphysics that will have the crucial consequence of greatly winnowing the sorts of explanation that are appropriate for my investigation of causation. So, I caution readers again not to be hasty in thinking they have fully grasped the essence of empirical analysis based on what I have stated so far.

1.2.1 EFFECTIVE STRATEGIES

A mundane but instructive observation about causation is that it generalizes a wide variety of other concepts like digestion, photosynthesis, rusting, erosion, gravitation, and combustion. One might say these are all species of causation. For each one, there exist conditions or events that are reliably connected to other conditions or events. Where there is abundant dry wood, plenty of oxygen, and a small fire, there will often be a larger fire shortly afterwards. Moreover, such causal regularities are largely insensitive to many other events. There are no remarkable relations between fires and the remote existence of goats or dirt or boron. The lack of notable connections between fire and so many other conditions is partly what makes fuel and oxygen noteworthy vis-à-vis fire.

My particular empirical analysis of the metaphysics of causation attempts to unify our understanding of such connections by concentrating initially on the following entirely unoriginal seed of an idea: *causes are means for bringing about certain kinds of effects*. The label drawn from Nancy Cartwright (1979) for this focus of causal talk is 'effective strategies'. It is empirically verifiable that combining dry wood, oxygen, and some source of heat is a good strategy for creating fire, whereas dunking an ordinary rock in water is demonstrably an ineffective method for starting a fire. The empirical analysis provided by my theory is aimed at facilitating explanations of why there is a regular pattern of events demonstrating that some strategies for affecting the world are better than others. 'Effective strategies' is the name for this empirical content of causation. The effectiveness of a strategy is testable (to a first-order approximation) simply by acting on the strategy a bunch of times, acting on alternative strategies a bunch of times, and observing whether the desired effect occurs more often after using the designated strategy than after using the alternatives.

It will take some work to unpack what 'effective strategies' ultimately amounts to and to ensure that the resulting empirical analysis makes sense of causation that does not involve strategies. This work will not be completed until chapter 5, but I can make a few preliminary comments here.

The attention placed on effective strategies is merely an educated guess about where to begin an exploration of the phenomena we pre-theoretically associate with causation. Nothing about this choice forecloses the possibility that other phenomena associated with causation can be incorporated or prevents an entirely different starting point from leading to a fruitful empirical analysis. Furthermore, nothing ensures that the totality of empirical phenomena relevant to the metaphysics of causation will form a cohesive collection in the end. It might turn out to be conceptually optimal to segregate the empirical phenomena into multiple distinct clusters having little to do with one another. Whether we should explain the empirical phenomena as a cohesive unit or instead as a patchwork of distinct groups of phenomena is not a matter to be decided in advance. What we can say initially is that it makes sense to investigate phenomena that appear to make sense of why we have causal terminology, and that seems to me to be captured in large part by the principle that some happenings are effective at bringing about other happenings of a certain kind. So long as this conception of the empirical focus of causal talk fits comfortably within a suitably broad construal of the empirical phenomena relevant to the metaphysics of causation—including causal regularities having nothing to do with agency—it will not matter that this preliminary choice is somewhat arbitrary.

A key point to keep in mind is that 'effective strategies' is not an expression to which I am attributing any technical meaning. It merely stands for the pretheoretical idea that some strategies for achieving desired goals are reliably better than others. So one important constraint on the content of 'effective strategies' is that it not take strategies too seriously metaphysically. Naïvely speaking, for a strategy to exist, there needs to be some agent reasoning about how to accomplish a goal, but for the purpose of explaining causation we want to avoid assuming that for causation to exist, there needs to be agency somewhere in the universe. Similarly, we also need our resulting investigation of causation to accommodate the existence of intermediate or borderline cases of agency in a way that exhibits graceful degradation.

A delightful ambiguity in the expression 'effective strategies' is that it suggests enough of a difference between accidental and law-like regularities to substantiate our conviction that causation is more than mere happenstance while not insisting that the reliable effectiveness of some strategies requires some empirically inaccessible non-accidentality. On the one hand, there can be accidental regularities that should not count as causal. On the other hand, if we assume from the beginning that the facts to be explained are precisely the set of non-accidental regularities, that would raise the question of how we could know whether a regularity is accidental or not. We would no longer have uncontroversially empirical phenomena as explananda. A salutary feature of empirical analysis is its compatibility with a flexible distinction between what is accidental and what is enforced by law, avoiding both extremes in an account of the empirical content of causation. This flexibility does not prevent us from invoking a distinction between law-like and accidental in our explanation of the empirical content; it just avoids *requiring* the distinction in order to make sense of the empirical content. To illustrate by analogy, a biologist should not adopt the task of explaining why creatures with souls behave intelligently but instead why creatures that seem to behave intelligently are able to. An explanation for intelligent behavior might postulate a soul, but to assume the soul in the first place would leave unclear whose behavior requires explanation.

'Effective strategies' suggests a flexible distinction between law-like and accidental by encouraging us to think of situations where an agent is selecting or controlling circumstances in order to bring about some desired effect. If a regularity holds even when an agent tests it in numerous circumstances, that potentially counts as strong evidence that the regularity is suitably law-like. No number of test situations will ever ensure that the regularity is law-like, but the fact that we are able to manipulate the world in order to test regularities means that if we account for a pattern of regularities in circumstances where people are trying to test them for potential violations, we in effect account for why there exists a pattern of regularities that look as if they hold by virtue of laws.

One collection of phenomena subsumed under the umbrella of 'effective strategies' is that across a wide range of different kinds of events, materials, and circumstances, there exist exploitable regularities where one type of event C is a good means for bringing about an event of type E. But there are also important *general* features of such regularities. Two in particular stand out.

The first is that it is seemingly impossible to exert influence in one direction of time to an event and then back in the opposite direction of time to another event. These are ***backtracking nomic connections*** because they first go in one temporal direction and then backtrack in the opposite temporal direction. (Warning: many philosophers use 'backtracking' misleadingly to refer only to the past-directed half of the backtracking.) The reason someone might hypothesize that nomic connections could backtrack is that frequently the occurrence of one kind of event is correlated with the occurrence of two kinds of effects in its future. Throwing a rock into a pond leads lawfully to a distinctive kerplunk sound and expanding ripples. One might wonder why it is not possible to increase the chance of ripples by making a kerplunk sound. Any such strategy, I think, is demonstrably ineffective except to the extent it exploits a future-directed strategy such as tossing a rock in the pond. ***Causal directness*** is this (seemingly correct) principle that a backtracking nomic connection between two events never does anything beyond what it already does by virtue of temporally direct nomic connections.

The second (and closely related) general feature of effective strategies is that there are apparently no effective strategies for influencing the past in useful ways. The empirical phenomenon associated with this claim can be roughly characterized as follows. People who are assigned the task of bringing about some future outcome—like writing a haiku or baking bread or establishing a viable human colony on Pluto—are sometimes able to accomplish that task at a significantly higher rate than people who are trying to avoid having that kind of outcome occur. But people who are assigned the task of bringing about some past event of type E—no matter what E is—never do any better or any worse (on the whole) at having an instance of E occur than people who are trying to prevent instances of E. Call this phenomenon the ***asymmetry of advancement***. We can advance some of our goals for the future but never our goals for the past.

The broad aim of my metaphysics of causation is to provide a conceptual structure optimized for explaining the empirical phenomena associated with effective strategies, understood broadly to include causal relationships that involve no agency. Along the way, in §5.8, I will provide a skeletal explanation of why effective strategies exist across a wide range of activities and an explanation of why causal directness holds and why there is an asymmetry of advancement. There are certainly other features of effective strategies that I will consider, but the upshot of an empirical analysis of the metaphysics of causation is as follows. If an optimal, or at least adequate, set of concepts can be developed that help to explain (in a "complete story" sense of explanation) the empirical phenomena behind effective strategies (broadly construed), then the *metaphysics* of causation will be largely solved. All there is to understanding the metaphysics of causation in the sense relevant to empirical analysis is just understanding how these empirical phenomena are related to fundamental reality. As I will discuss in chapters 5 and 10, no one is currently in a good position to provide an explanation of all the details related to effective strategies, and no one is currently in a good position to provide an adequate comprehensive theory of fundamental physics (much less, fundamental reality). But the *empirical analysis* of the metaphysics of causation does not require that we explain everything about effective strategies; it merely requires a justification of the conceptual architecture that connects the empirical phenomena to fundamental reality. I hope readers will judge that the system of concepts that I will soon introduce are flexible enough to be applicable to a wide range of ways fundamental reality could be structured and to be applicable to any causal regularity, but also inflexible enough to facilitate non-trivial empirical predictions.

1.3 Empirical Analysis of the Non-metaphysical Aspects of Causation

My empirical analysis of the metaphysics of causation addresses causation insofar as we want to tailor our understanding of causation to structures "out there in reality" that are not very closely tied to how we think about causation. But there is a second empirical analysis that is adapted to address further empirical phenomena concerning how we conceive of causation, including causation's role in explanation and the discovery of causal regularities. I will refer to this second investigation as the ***empirical analysis of the non-metaphysical aspects of causation***. The purpose of this section is to clarify how the general methodology of empirical analysis applies to the aspects of causation that go beyond the scope of metaphysics, as precisified in this chapter. Its prominent components include (1) the psychology of causation, (2) the role of particular (or token) causes in the explanatory practices of the special sciences, and (3) causal modeling that is sufficiently remote from (or insulated from) the character of fundamental reality.

First, the subfield of psychology dedicated to exploring how people think of causation produces models that attempt to explain uncontroversially empirical data, including people's reactions when they are told stories or shown a sequence of events and asked, "What do you think caused this event?" or "Does event *c* count as one of the causes of event *e*?" This psychology of causation is also meant to be compatible with related phenomena such as how long children look at certain temporal sequences designed to mimic or violate default rules of object behavior, how people conceive of the operation of gadgets, how people attempt to solve mechanical puzzles, etc. Like other scientific theories, these psychological theories can be formulated using a technical vocabulary, distinguishing foreground and background causes, proximate and distal causes, actual and potential causes, etc.

One result we should expect from such a psychologically oriented empirical analysis is that its structures will almost certainly be significantly different from the structures of an empirical analysis aimed at explaining effective strategies. This is easy enough to motivate by virtue of the general pattern whereby a theory of *X* often looks very different from a theory of the psychology of *X*. A scientific theory of space, for example, is aided by having esoteric mathematical structures like manifolds and curvature tensors; a scientific model of how humans naturally think about spatial relations is sure to exclude curvature tensors in favor of structures that better represent the portion of our cognitive processing that manipulates spatial information.

We know enough about our psychology to recognize that humans use various heuristics to understand causal connections in the external world in a simplified way. Because people are poor reasoners about fantastically small probabilities, we should expect them to oversimplify causal relations that involve minute probabilities. People have limited capacity in their working memory, so we should expect them to ignore some causes when a vast multitude of causes are present. There is no a priori reason why the scientific conception of causation *must* differ from our implicit pre-theoretical conception of causation, but it should not be even remotely surprising. More important, there is no reason to assume from the outset that there must be some interesting causation concept simultaneously optimized for both explaining the core phenomena behind causation itself (as some relation out there in reality) and explaining regularities concerning our instinctive causal judgments.

Second, although philosophers do not ordinarily consider causal explanation a topic in psychology, I will discuss in chapter 8 a sense in which disputes about causal explanation—over which individual events explain a particular effect—are psychologically oriented to the extent that they go beyond the "complete story" explanations afforded by the totality of causal relations in my metaphysics. If two people agree on the fully detailed account of how the effect *e* came about by agreeing on all the relevant laws and how they connect the complete arrangement of every last bit of matter, then any further disagreement about which partial causes are *explanatory* cannot be adjudicated by reference to further empirical data about the events leading up to *e* because there would be no further empirical data. The only extent to which such a debate can be informed by reference to further empirical data would come from investigations of people's explanatory practices, including their revealed explanatory norms. In this limited sense, empirical analysis treats causal explanation like it treats the psychology of causation.

Third, there are many invocations of causation in the special sciences, especially the practice of discerning causal relations from statistical correlations. This includes the scientific and philosophical literature on causal modeling (Spirtes, Glymour, and Scheines 2000; Pearl 2000; Woodward 2003). Much of this scientific activity can be understood without any particular connection being drawn to fundamental reality. As such, these investigations of causation do not count as metaphysical in the framework I have adopted; they count as part of the special sciences, to be addressed by an empirical analysis of the non-metaphysical aspects of causation.

Finally, we should expect an empirical analysis of the metaphysics of causation and an empirical analysis of the non-metaphysical aspects of causation to be related in a fairly straightforward way. The reason people have a concept of causation is that it provides an efficient way to conceptualize those structures responsible for things behaving causally. The metaphysics of causation directly addresses why things "out there in reality" behave causally, why some kinds of events reliably bring about certain other kinds of events. The other empirical analysis addresses how the structures posited in the metaphysics might be simplified for cognitive consumption, paying special attention to people's need to learn about effective strategies and apply them to new circumstances. We ought to suspect our psychology of causation will match, to a first order approximation, the structures that ultimately account for the world behaving causally, but a second order correction would likely take into account our need for an efficient cognitive grasp of these structures. We also ought to suspect that our judgments about which causes are explanatory will fit into the general cognitive system that filters the vast plenitude of partial causes for events that have some cognitive salience. Expressed curtly, we have intuitions and practices for identifying certain partial causes as explanatory as a by-product of their role in our cognition, especially by virtue of our heuristics for learning about effective strategies.

Although I will sketch a theory along these lines in chapter 9, even these suspected connections are not inviolable constraints on the psychology of causation or our theories of causal explanation because the structures that explain the empirical phenomena associated with effective strategies might be so complicated or so remote from our epistemic access to reality that our cognition only makes contact with the metaphysics of causation through roundabout means.

To summarize these last two sections, the application of empirical analysis to causation results in two scientific investigations. The first explores the empirical phenomena related to causation as something "out there in reality," what ultimately becomes the metaphysics of causation. The second explores further aspects of causation that are based on how creatures think about causation. This bifurcation is analogous to how a scientific investigation of food can be divided between an investigation of the nutritional aspects of food and an investigation of food that goes beyond its nutritional aspects to its social role and our usage of the word 'food'. Because the primary goal for this volume is to provide a scientific metaphysics of causation, I will only discuss the empirical analysis of the non-metaphysical aspects of causation in order to show how it relates to the metaphysics of causation and to illustrate how some traditional philosophical problems concerning causation can be resolved when they are properly situated outside the scope of metaphysics.

1.4 Causation as Conceptually Tripartite

Now it is finally time to turn our attention to the structure of my theory of causation. I will initiate the discussion by explaining how the concept of causation should be divided into three stacked layers: bottom, middle, and top. The bottom and middle layers are relevant to the metaphysics of causation whereas the top layer pertains to the non-metaphysical aspects of causation. Then, I will draw a distinction between fundamental reality and derivative reality and describe how the bottom conceptual layer of causation concerns fundamental reality whereas the middle and top layers concern derivative reality. Last, I will draw a distinction between two different sets of standards for evaluating theoretical adequacy, STRICT and RELAXED, and I will defend the thesis that one's metaphysics of causation, the bottom and middle layers, should be evaluated according to STRICT standards whereas an empirical analysis of the non-metaphysical aspects of causation, the top layer, can be entirely adequate even if it only satisfies the more permissive RELAXED standards.

Philosophers who have weighed in with positive theoretical accounts of causation have often focused on a single proffered core aspect of causation—determination by the laws of nature (Mackie 1973), counterfactual dependence (Lewis 1973b), probability-raising (Suppes 1970), transference of some privileged sort of physical quantity (Salmon 1977, Kistler 1999, Dowe 2000)—and have tried to show how legitimate causal claims are vindicated primarily in terms of that one core aspect, whether they involve a magnetic field causing an electron to accelerate or an increase in literacy causing a redistribution of political power. Other theories (Good 1961, 1962, Sober 1985, Eells 1991, Salmon 1993, Hall 2004) are conceptually dual in the sense that they try to make sense of causation in terms of two core causal concepts that operate mostly independently of one another. My own account models causation in terms of three distinct but related conceptual layers.

That my analysis segregates the concepts it constructs for understanding causation into three layers rather than one, two, or forty-seven is not in itself particularly noteworthy. There is no prima facie reason to expect a three-layer account to be superior to a dual or quadripartite account. The tripartite decomposition is merely the result of a natural division of labor concerning what a theory of causation should rightly be expected to accomplish. The two principles that divide the three conceptual layers are these:

1. There is an important metaphysical distinction between that which exists fundamentally and that which does not.
2. There is an important methodological distinction between how one should evaluate a theory of the metaphysics of causation and a theory of the non-metaphysical aspects of causation.

My account of causation thus divides the concepts it employs into three layers corresponding to (1) those appropriate to fundamental reality, (2) those appropriate to derivative reality insofar as it bears on the empirical phenomena associated with the metaphysics of causation, and (3) those appropriate to derivative reality insofar as it bears on the empirical phenomena associated with the non-metaphysical aspects of causation. No single layer, by itself, contains a relation that deserves to be designated as *the* causal relation, but all three layers together constitute a collection of concepts that allow us to make adequate sense of everything regarding causation that needs to be accounted for.

The consequences of this tripartite division are significant and set apart my theory from other existing accounts. Most accounts of causation maintain that there is a cause-effect relation between individual chunks of reality with certain distinctive characteristics. For one thing, the postulated cause-effect relation often holds among mundane objects or events or facts,[3] like a cloud casting a shadow or a virus provoking an immune response, rather than holding only between spatially expansive and microphysically detailed states. For another, the cause-effect relation is normally taken to be irreflexive because it is believed that effects do not cause themselves. Finally, the cause-effect relation is also thought to be non-symmetric because effects do not cause their causes, except perhaps in special circumstances like a time travel scenario. On my account, this crude cause-effect relation has no place in the metaphysics of causation but is suitable for the top layer where the epistemological and psychological roles of causation are properly situated. Relocating the cause-effect relation out of the metaphysics serves to dissolve a large number of problems philosophers routinely assume need to be resolved decisively by any adequate account of causation.

[3] Existing theories vary greatly in their metaphysical account of the causal relata, e.g. whether they are events, property instantiations, aspects, processes, tropes, etc. They also vary in whether they include additional parameters. Causation might not just be a two-place relation between the cause and effect, but a three-place or four-place relation, where the extra parameters can be contrasts, processes, choice of causal variables or choice of causal model, etc. Despite all such differences, most existing accounts of causation are such that when all the additional parameters are filled in, the residual relation between cause and effect shares much of the logical character of folk attributions of causation.

TABLE 1.1 The Three Conceptual Layers of Causation.

Layer	Subject	Metaphysical status	Standards of adequacy
Top	Non-metaphysical aspects	Derivative	RELAXED
Middle	Derivative metaphysics	Derivative	STRICT
Bottom	Fundamental metaphysics	Fundamental	STRICT

The three layers are depicted in Table 1.1. One prominent difference among the layers concerns whether they apply to singular or general causation. *Singular* causation applies to cause-effect relations that occur in a single fragment of the world's history. Examples of singular causal claims include, "The collapse of the Tacoma Narrows Bridge was caused by wind," and "The cholera outbreak was caused by a contaminated well." *General* causation addresses the *kinds* of events that can cause some chosen *kind* of effect. Examples of general causal claims include, "Smoking causes cancer," and "Bribes encourage corruption."

The bottom layer addresses an extremely inclusive form of ***singular causation*** in terms of a theory of fundamental reality employing concepts like determination and probability-fixing. The middle layer addresses ***general causation*** in a way that abstracts away from the details of fundamental reality. The top layer addresses the less inclusive form of singular causation that people employ in everyday conversations and that scientists employ when giving causal explanations of particular effects. I call these singular causes 'culpable causes.'

These three conceptual layers exhaust the scope of my account of causation. The overall structure of this book—after the methodological issues have been dealt with in this chapter—is simply to fill in the details concerning the bottom, middle, and top layers. The concepts in each layer depend on the resources of the layers underneath, so it is wise to consider the layers from the bottom up.

My goal for the rest of this chapter is to demarcate the three conceptual layers in greater detail. First, I will provide a simplistic overview of my account of causation. Next, I will elaborate on the distinction between fundamental and derivative in order to make clear how the bottom conceptual layer of causation differs from the two layers above it. After that, I will unpack my distinction between STRICT and RELAXED standards of theoretical adequacy, which separates the top layer from the two layers beneath it. Finally, at the end of this chapter, I will return to the three conceptual layers of causation in order to recap how they relate to singular and general causation.

1.5 A Sketch of the Metaphysics of Causation

Before I explicate the more idiosyncratic elements of my overall account, it will likely be useful for me to summarize simplistically how my account will eventually help explain how causes are effective at bringing about their effects. To keep the discussion manageable, I invite readers to accept provisionally that there are some fundamental laws of physics that govern the behavior of all particles and fields and that ordinary macroscopic objects are merely aggregates of these fundamental microscopic parts.

Imagine a magnetic compass lying undisturbed. By moving a lodestone near the compass, one can reliably make the compass needle move. It is uncontroversial and empirically verifiable that events of type C, moving a lodestone near a compass, are effective at bringing about events of type E, a jostling of the needle. What explains such phenomena, according to my account, is that there exists a fundamental reality that includes extremely detailed facts about how fundamental particles and fields are arranged as well as fundamental laws governing the temporal development of this fundamental stuff. The objective structure behind all causation is located in how the fundamental laws link the fundamental material stuff at different times and places. Specifically, some fundamental happenings *determine* the existence of other fundamental happenings or *fix an objective probability* for their existence, and that is what ultimately grounds all causal relations.

Yet, when we explore the character of plausible fundamental laws, we find good reason to believe that fundamental laws by themselves provide no connection between the highly localized event, c, constituting a particular lodestone's motion toward the compass, and the consequent jostling of the needle, e. At best, the fundamental laws connect e only with a much larger event c' that includes c as well as a complete collection of microphysical facts occurring at the time of c and occupying a vast expanse of space, perhaps stretching out to infinity. Puny events like c are too small to determine or fix any objective probabilities for events like e, but gargantuan events like c' can. The localized chunk of reality c only plays a fundamental causal role by virtue of its being a part of the much larger c'.

That story is fine as far as fundamental reality goes, but because we humans are unable to perceive microphysical states accurately enough, unable to reckon their nomic consequences accurately enough, and unable to control the world precisely enough, these fundamental relations are by themselves rarely useful to us in practice. Fortunately, our world is amenable to various approximations that allow us to represent aspects of fundamental reality in ways that abstract away from their precise character. Our belief that c caused e is in part a belief that the lodestone part of the world was somehow a more important part of the vast c' than all the far flung events that seemingly have nothing to do with the motion of the compass needle. What makes c the important part of c', I claim, is that the probability that c' fixes for the effect is significantly greater than the probability that would be fixed for the effect by events that are just like c' except that the physics instantiating the movement of the lodestone is hypothetically altered to make the lodestone remain at rest. The motion of the lodestone is causally important to the compass needle because it affects the needle's probability of moving.

The metaphysical picture, boiled down to its essence, is that there is some sort of fundamental reality instantiating relations of determination or probability-fixing among microscopically detailed events and a more abstract or fuzzy construal of reality where events of type C raise the probability of events of type E. This helps to explain the existence of effective strategies because, to a first approximation, an effective strategy for bringing about some instance of E is to bring about an event that raises the probability of E. My task in the rest of this book is to provide the resources to facilitate talk of probability-raising, determination, influence, etc. so that the details of this explanation can be specified in an acceptable way.

1.6 Fundamental and Derivative

The account of causation I present in this book crucially relies on a metaphysical distinction between fundamental and derivative. Most people, I think, have at least some intuitive grasp of the difference between fundamental and derivative, and for the purpose of understanding causation, we can mostly rely on that intuitive grasp. However, in order to focus the distinction a bit more, I will list a few guiding principles and then describe an example of how we can think of kinetic and thermal and mechanical energy as derivative properties that reduce (in a sense I will soon clarify) to fundamental attributes, for example, mass and relative speed. (Throughout this book, an ***attribute*** is a property or relation, broadly construed.) This example will serve as a template for my account of causation, clarifying how derivative aspects of causation can be related to fundamental aspects of causation.

I will attempt to characterize fundamental and derivative reality without introducing undue controversy, but because the distinction bears on broad principles of ontology, truth, and explanation, there will inevitably be plenty of room for disagreement about its ramifications. Because even my best effort to make precise my intended conception of fundamentality will suggest a conclusion that is objectionable to some readers, I want to emphasize from the start that for the purpose of applying the distinction between fundamental and derivative to causation, not everything I say about fundamentality is absolutely essential. My goal here is merely to establish an initial reference. Readers who disagree with me on a few points here and there should be able to grasp the gist of my distinction and translate it into their preferred terminology.

I think the easiest way to get a grip on what is fundamental and what is derivative is to start by thinking about reality in a rather naïve way as *existence*. Just consider everything that exists, including all objects, properties, relations, substances, and whatever else you think needs to be included. The totality of existents, including all their relations with one another, constitutes reality. Then, we can think of reality as subdivided into exactly two disjoint parts: fundamental and derivative. 'Fundamental' and 'derivative' at this point serve as placeholders for a distinction that is filled in by specifying the role that 'fundamental' plays.

I am now just going to present a list of some platitudes that appear to me to capture several constitutive features of fundamental reality:

1. The way things are fundamentally is the way things *really* are.
2. Fundamental reality is the only real basis for how things stand derivatively.
3. Fundamental reality is as determinate as reality ever gets.
4. Fundamental reality is consistent.

While these principles are admittedly vague and subject to philosophical objection, I think they provide a useful starting point for discussing fundamental reality.[4] I will attempt to specify them more precisely by laying out a specific example based on the concept of kinetic energy, which will serve as a reference for further clarification.

For our purposes, it will be helpful to simplify by operating provisionally under the pretense that every existent is either determinately fundamental or determinately derivative. After the basic distinction is clear enough, one can take up the project of evaluating the extent to which there is indeterminacy at the boundary between fundamental and derivative reality.

1.6.1 THE KINETIC ENERGY EXAMPLE

The theory of classical mechanics is a scheme for modeling how material bodies move around according to forces. I will focus on a specific interpretation of classical mechanics whose purpose is to clarify ontological commitments: ***the simple theory of classical mechanics***. There are other ways to interpret the content of classical mechanics, but I am not engaging here with technical issues in the philosophy of physics or with historical exposition.

The ingredients of the simple theory of classical mechanics include a space-time inhabited by corpuscles bearing intrinsic properties like mass and charge. A ***corpuscle*** is by definition a point particle, meaning that it has an identity through time and occupies a single point of space at any given moment so that its history over any span of time is a path in space-time. Corpuscles in classical mechanics rattle around according to exceptionless laws where each corpuscle's acceleration is a relatively simple mathematical function of fundamental attributes like the inverse-square law of gravity and some sort of short-range repulsive interaction that makes corpuscles bounce elastically away from each other when they (nearly) collide. To be more specific, the simple theory posits the following structures: an appropriate space-time, corpuscles, charge and mass properties that adhere to the corpuscles, a distance relation between any two corpuscles at any given time, a relative speed relation between any two corpuscles at any given time, and a law

[4] I provide further discussion of these principles and their role in Empirical Fundamentalism in Kutach (2011b).

governing how these attributes evolve over time. The simple theory posits nothing else. Once all these entities and attributes have been everywhere specified, the entire world has been specified according to the simple theory.

We know that classical mechanics is not an accurate theory of our world, but for pedagogical purposes, it is convenient to consider how we ought to think about reality if it were true that the actual world perfectly matched one of the models of the simple theory of classical mechanics. For the rest of this section, discussion will proceed under the pretense that some model of the simple theory of classical mechanics is the complete and correct account of fundamental reality so that we have a concrete reference for understanding fundamentality. Having adopted the simple theory of classical mechanics, we can distinguish between fundamental and derivative in a fairly intuitive way. The corpuscles and space-time are fundamental entities, their relative distances and speeds are fundamental relations, their masses and charges are fundamental properties, and the laws governing them are fundamental laws. They are all fundamental existents. Poetry, patience, and financial assets, by contrast, are arguably non-fundamental. They do not appear as components or parts of the model nor do the laws of the simple theory make special use of them. It is uncontroversial that poetry, patience, and financial assets exist. Therefore, assuming they are not fundamental existents, they are derivative existents.

In more generality, once we have supposed that some model completely and accurately represents fundamental reality, we can think of derivative entities and properties as existents that are not substructures of the model.

One important group of existents whose status deserves to be considered for illustration are "moderate sized specimens of dry goods" (Austin 1961). Without getting too bogged down in technicalities, I think the most natural method for categorizing macroscopic material objects can be sketched as follows. If fundamental reality includes a space-time containing the corpuscles and fields that instantiate an ordinary material object, then any particular *instance* of this object—that is, a full specification of the complete microscopic content of a maximally determinate region of space-time that includes at least one temporal stage of the object—will be a part of fundamental reality and thus will be a fundamental existent. But insofar as we treat an object as an existent that retains its identity under even the slightest alterations to its boundary or its microscopic instantiation, or its time or place of occurrence, we are treating it as a derivative existent. (There is no fact of the matter in this case as to whether the *object* is fundamental or derivative. There are instances of the object, which are all fundamental, and there are various abstractions or fuzzings or coarse-grainings of the object, which are all derivative.) Alternatively, if fundamental reality is something more esoteric like an entangled quantum field or an eleven-dimensional arena inhabited by strings, then there may well be no parts of fundamental reality that count as an instance of the object, in which case the object is unambiguously derivative.

Insofar as discussion in this volume will be concerned, it will suffice for us to adopt a single sufficient condition for an existent being derivative. I will specify this sufficient condition in terms of the ontological status of *quantities*, but it holds of existents generally.

> A quantity is derivative if its magnitude requires some specification beyond the totality of fundamental reality (and beyond any specification required to locate the quantity in fundamental reality).

For example, the mass of any corpuscle at any time in the simple theory of classical mechanics is a quantity that has a determinate magnitude once we have specified the spatio-temporal location of the corpuscle whose mass we are considering. By contrast, the kinetic energy of any corpuscle (at any time) is derivative. A corpuscle's ***kinetic energy*** is equal to one-half its mass times its speed squared, $\frac{1}{2}mv^2$. In the simple theory of classical mechanics, no fundamental structures suffice for a given corpuscle's absolute speed; there are only corpuscle speeds relative to other corpuscles. However, if we select an appropriate frame of reference to serve as a universal standard for being at rest, we can say that a corpuscle's speed is its speed relative to this rest frame. Then, because we have associated a determinate speed with each corpuscle, there will be a well-defined value for each corpuscle's kinetic energy. The kinetic energy of a corpuscle is an example of a derivative quantity because there is nothing in fundamental reality that corresponds to a unique correct value for the kinetic energy (at the corpuscle's spatio-temporal location) unless we augment the model with a parameter that doesn't correspond to anything in fundamental reality, namely this stipulation of what counts as being at rest.

Whenever a parameter used for describing reality does not have a unique correct assignment given how fundamental reality is structured, let us say that it is ***fundamentally arbitrary***. A choice of rest is one example of a fundamentally arbitrary parameter. More generally, coordinate systems and so-called gauge degrees of freedom are fundamentally arbitrary. Fundamental reality might make some coordinate systems more convenient than others for characterizing the distribution of matter, but fundamental reality itself is independent of our conventions for assigning labels to points of space-time. Any quantity that is coordinate-dependent is derivative.

By convention, we can adopt the policy that the fundamentally arbitrary specification needed to locate a region in space-time (or whatever space is the container of fundamental material stuff) does not by itself make the contents of that region derivative. The locating information should instead be interpreted as merely defining the component of fundamental reality under consideration.

Imagine two solid blocks in an otherwise empty portion of space, each composed of massive corpuscles bound together by short-range forces. Fig. 1.1 provides two different characterizations of the very same fundamental arrangement of corpuscles that constitute the two blocks. By choosing a rest frame, one bestows on each corpuscle a well-defined (non-relational) velocity. The total kinetic

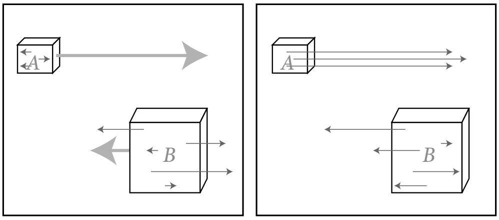

FIGURE 1.1 *Two depictions of the same fundamental reality. On the left, thermal energy is calculated by treating a corpuscle's speed relative to the motion of its block. On the right, thermal energy is calculated by treating every corpuscle's speed as relative to the rest frame.*

energy, E_T, of the entire system is the sum of each individual $\frac{1}{2}m_i|\vec{v}_i|^2$, where m_i is the mass of the ith corpuscle and $\vec{v}_i$ is the velocity of the ith corpuscle in the rest frame. We can think of this total kinetic energy as divisible into ***thermal energy***, and ***mechanical energy***, with the relative proportion depending on how we choose to organize the complete collection of corpuscles into groups. Thermal and mechanical energy are both forms of kinetic energy, at least insofar as we are simplifying the physics in this volume for the sake of discussion.

One way to group the corpuscles is to let the A-grouping comprise all the corpuscles in the block marked A, and the B-grouping comprise all the corpuscles in block B. Let $\vec{v}_A$ be the velocity of block A and m_A be the mass of block A, and similarly for $\vec{v}_B$ and m_B. In Fig. 1.1, these two velocities are represented as large gray arrows. The corresponding mechanical energy of block A is $\frac{1}{2}m_A|\vec{v}_A|^2$, and the mechanical energy of block B is $\frac{1}{2}m_B|\vec{v}_B|^2$. The thermal energy of each individual corpuscle can be understood as its kinetic energy *relative to the net motion of its group*. Each corpuscle i on the A-grouping, for example, has thermal energy $\frac{1}{2}m_i|\vec{v}_A - \vec{v}_i|^2$. The thermal energy of block A as a whole is just the sum of the thermal energy of each of its individual members and similarly for block B. The total thermal energy of the whole system is just the sum of the thermal energy of each block.

A second way to group the corpuscles is to put all of them together. Let m_T be the total mass of all the corpuscles and $\vec{v}_T$ be the velocity of the center of mass of the whole system. Then the mechanical energy of the whole system is $\frac{1}{2}m_T|\vec{v}_T|^2$, and the thermal energy is the sum of all the individual terms of the form $\frac{1}{2}m_i|\vec{v}_T - \vec{v}_i|^2$.

The first decomposition of kinetic energy into mechanical and thermal quantities fits our natural inclination to treat the blocks as separate objects and is useful for making predictions about thermodynamic phenomena when each block comes into thermal contact with other objects having their own distinct temperature and composition. If block A is grabbed and put in contact with some ice, and block B is separately placed in a furnace, the thermal energies as calculated in the first decomposition are what figure in predictions of how heat will move between the blocks and their respective environments.

The second decomposition of the kinetic energy into mechanical and thermal quantities is not useful for such calculations. However, there is nothing formally incorrect about it, and there might be some circumstances where that way of breaking apart the kinetic energy into thermal and mechanical components is more useful. The important point here is just that nothing about fundamental reality makes one assignment of mechanical and thermal energy the unique correct assignment, even if fundamental reality makes one assignment especially useful for practical purposes. The allocation of kinetic energy between mechanical and thermal energy is thus fundamentally arbitrary.

Furthermore, there are circumstances where there is no clear best way to distinguish between thermal and mechanical energy even for practical purposes. The oceans have mechanical energy in their currents and thermal energy that plays a role in melting icebergs, but because the ocean is fluid, it can sometimes be unclear how to group the corpuscles. The tiniest eddies in the current might be construed as instantiating purely mechanical energy because they knock pollen grains around, but they could be construed as purely thermal because such a small parcel of energy cannot be extracted by any practical device like a turbine.

Now we are in a position to see why it is reasonable to think of mechanical and thermal energy as metaphysically derivative. First, we know that mechanical and thermal energy can be defined in terms of properties we already accept as fundamental by specifying two fundamentally arbitrary parameters, the choice of rest and the grouping of corpuscles. That leaves us with a choice about how to construe the amounts of thermal energy and mechanical energy:

- One option is to hypothesize that there is a brute fact of the matter about precisely how much thermal and mechanical energy the system has. I interpret such a choice as an addition to what we already accept as part of fundamental reality. This is tantamount to believing there are objective facts about the distribution of thermal and mechanical energy that go beyond how the fundamental masses are arranged in space-time, and a good way to describe such facts is to say that they are fundamental.
- A second option is to declare that there is no ultimate fact of the matter about how much thermal and mechanical energy there is, but that there are still parameter-dependent facts about thermal and mechanical energy. Given that the corpuscles are arranged in such and such a pattern fundamentally, and given that we choose such and such frame of reference for the rest frame, and given that we allocate these corpuscles over here to the A-group and those over there to the B-group, there is a determinate value for both the thermal and mechanical energy of A and B. A good way to describe this option is that it treats thermal and mechanical energy as derivative.

There are other interpretational options one could consider, but I will forgo discussion of them because my goal here is just to provide a reference point for how to distinguish fundamental from derivative, not to settle quibbles about how best to understand energy.

I believe the more reasonable stance is to interpret kinetic energy (and thus thermal and mechanical energy) as derivative. There are several reasons to prefer treating them as derivative rather than fundamental. (For concision, I will focus attention on kinetic energy in this paragraph, but everything I say here applies to thermal energy and mechanical energy as special cases.) First, we already have fundamental laws in classical mechanics governing the motions of particles, and if there were some brute (fundamental) fact about precisely how much kinetic energy existed, it would play no essential role in the temporal development of the physics.[5] Second, if there were a fundamental fact about the precise quantity of kinetic energy, we would have no epistemic access to its value. At least, it would be mysterious how we could ever come to know the one true amount of kinetic energy. Third, there is no scientific account of anything that would be defective in any way if we treated kinetic energy as derivative. Nor would any scientific account be improved by treating it as fundamental. These kinds of considerations are standard in scientific practice and provide a practical grip on why we construe some quantities as fundamental and others as derivative. A good way to think about the issue is that if we try to segregate various kinds of properties and relations into fundamental and derivative using scientific methods, we have good reasons to keep the fundamental ontology fairly restricted and to avoid postulating redundancies in fundamental reality. Ceteris paribus, a sparse theory of fundamental reality can provide more reductive explanations, posit fewer epistemically inaccessible facts, posit fewer quantities that do not integrate well with the rest of the fundamental quantities, etc. (A principle of parsimony could conceivably be included in the list of principles associated with the idea of fundamentality, but I think it is ultimately preferable to leave such scientific considerations out of the constitutive principles governing fundamentality in order to accommodate a broader range of approaches toward fundamental reality.) Although I have not argued conclusively that kinetic (and thus thermal and mechanical) energy should be accorded derivative status, the thesis has a lot to recommend it, so from here on, the discussion will assume they are to be understood as metaphysically derivative.

1.6.2 SOME CONSTITUTIVE PRINCIPLES OF FUNDAMENTALITY

With the kinetic energy example in mind, we can now revisit the list of principles I associated with fundamentality.

[5] It is possible to formulate classical mechanics so that energy plays a starring role in the temporal development, but the simple theory was constructed to exclude energy from any essential role in the fundamental laws. In any case, kinetic energy by itself plays no role in the fundamental development of the actual world even if energy itself does.

Principle (1) claims that the way things are fundamentally is the way things *really* are. This is a way of assigning privileged metaphysical or ontological status to fundamental reality. Because the topic of ontology is too controversial to take up here, I will only note that philosophical debates over realism and anti-realism can be usefully framed in terms of fundamental reality, and it is one of the main goals of Empirical Fundamentalism to reformulate debates about realism in terms of debates about what is fundamental.

For example, my suggestion that we should interpret kinetic energy as metaphysically derivative accords with Marc Lange's (2002) discussion. Lange cites the frame dependence of kinetic energy as a reason to believe that kinetic energy is not real, unlike a corpuscle's mass, which is well-defined independent of any choice of reference frame or coordinate system. If I have interpreted Lange correctly, what I mean by 'fundamental' is what Lange means by 'real'. This suggests that there is at least one construal of the word 'real' that tracks fundamental existence, though certainly our pedestrian attributions of 'real' apply to both fundamental and derivative existents.

Before going any further, I want to emphasize that a commitment to the existence of fundamental reality, as I construe it, does not impose significant constraints on a theory of reality. Although the simple theory of classical mechanics draws on a familiar conjecture that what is fundamental is the stuff described by theories of fundamental physics, nothing in my account of fundamentality or my account of causation requires that what is fundamental include physics, much less be identified with some instance of fundamental physics. As far as my theory is concerned, one could hold that economic events or geological processes or thoughts are fundamental. Because fundamental reality can in principle comprise just about anything, the mere claim that reality divides into fundamental and derivative has little substantive content. It can only generate rich consequences when conjoined with significant constraints governing fundamental reality.

For that reason, throughout much of this book, I will be exploring the auxiliary hypothesis that fundamental reality resembles models of paradigm theories of fundamental physics. This assumption is not strictly part of the theory of causation; its only purpose is to permit discussion of causation in a concrete context. It would be extremely difficult to say anything interesting about causation without at least exploring some auxiliary hypotheses about the nature of fundamental reality, and a focus on fundamental physics as a preliminary working model for understanding fundamental reality is motivated by the privileged role fundamental physics plays in any scientific investigation of causation that purports to hold for all kinds of causes. Physics has a distinguished role to play because a comprehensive theory of causation is supposed to apply not only to mundane affairs but also to the fantastically small and fantastically large, domains where only physics has provided a rich account of how things operate. Though one's theory of causation should apply to oceans and economies and psychological processes, it is at least a plausible hypothesis that the empirical phenomena that lead us to believe in causal relations described by the special sciences are instantiated by matter that obeys laws of physics. I take it as not even remotely plausible that causation among neutrinos and quarks could be cashed out in terms of oceanic, economic, or mental properties. But the reverse relation—that causation among economic events, say, is a special case of causation among the entities of fundamental physics—is at least a plausible working conjecture.

Although fundamental physics superficially says nothing about economics, it is easy to see skeletally how the laws of physics could impose extremely strong constraints on the physical stuff that instantiates paradigmatic economic activity. If the physical laws are deterministic, for example, a complete specification of the physical state at one time determines the physical arrangement of everything throughout history, including bankers at work, money in people's pockets, merchandise on the store shelves, and just about everything else that is economic in character. It may well be true that even a highly idealized epistemic agent cannot make successful economic predictions knowing the detailed physical state and the deterministic laws, and it is certainly true that the determination does not hold merely by virtue of the economic facts at one time, but it should be easy to understand how there could be non-trivial implications among the physical instances of economic facts.

Another good reason to investigate the auxiliary assumption that fundamental reality resembles existing paradigm theories of fundamental physics is that it is possible to derive remarkable facts about causation from a few relatively uncontroversial hypotheses about the fundamental physical laws. For example, in §6.2, I will derive causal directness, the principle from §*1.2.1* that a backtracking nomic connection does nothing beyond what it does by virtue of operating in a single direction of time. In chapter 7, I will do the same to demonstrate that the past cannot be usefully manipulated. Interestingly, none of my arguments defending these principles presupposes a fundamental asymmetry of causation or a fundamental passage of time or any settledness of the past. So, my explanation of the direction of causation will hold even if the fundamental laws are deterministic in both temporal directions with no fundamental temporal asymmetry.

Regardless of the benefits of my focus on fundamental physics as a guide to fundamental reality, the framework I will be constructing is suitable for a wide range of possible views about what should be included in fundamental reality. It is compatible with models of fundamental reality that include phenomenal properties, theological attributes, intentionality, and aesthetic properties. It is compatible with models of fundamental reality where nothing is physical. Emergent properties and dualistic conceptions of the mind can also be represented within the confines of the framework. Nothing I say about causation rules out any of these possibilities. I will just be using physics as a preliminary working model to help guide our thinking about causation.

Principle (2) claims that fundamental reality is the only real basis for how things stand derivatively. Philosophers have tried to make this idea precise in a variety of ways. One way is to say that fundamental reality fixes derivative reality. Another is to think of fundamental reality as a universal truthmaker, something by virtue of which all true claims about reality are true. Yet another option is to think that what is fundamental serves as a supervenience base for derivative reality. Although I believe such options are aimed approximately in the correct direction, I suspect all of these existing approaches will ultimately provide a suboptimal model for understanding the relationship between fundamental and derivative. To avoid unnecessary controversy, however, I will not provide my own account in this first volume of Empirical Fundamentalism. Instead, see Kutach (2011b) for a sketch of my views on this topic and future volumes for more details.

For present purposes, it will suffice to consider one critical feature of the kinetic energy example that needs to be accommodated by any account of how fundamental reality serves as the "real basis" for everything derivative. Remember that in order to derive any specific value for mechanical energy, say, one needs the fundamentally arbitrary choice of rest and corpuscle grouping. These parameters do not represent some additional fact about fundamental reality; they are stipulations. A complete specification of the fundamental attributes of classical mechanics does not by itself suffice for any particular value whatsoever for mechanical energy. How things are situated fundamentally does not fix how much mechanical energy there is. Yet, any specific choice of parameters will imply a precise amount of mechanical energy. So, given a complete characterization of fundamental reality, there exists a complete conditional characterization of mechanical energy, a complete set of conditionals of the form, "If such and such choice of rest and such and such choice of corpuscle groupings are made, the mechanical energy is such and such." Thus, how things are situated fundamentally (assuming the simple theory of classical mechanics) necessitates how things stand with regard to mechanical energy once (and only once) we have chosen the appropriate fundamentally arbitrary parameters.

Although the kinetic energy example shows how a numerically precise quantity can be conditionally implied by fundamental reality, my conception of derivative existents does not require such a conditional implication in order for them to count as bona fide existents.

Principle (3) claims that the way things are fundamentally is as determinate as reality ever gets. The fact that we have to supplement the fundamental attributes of classical mechanics with fundamentally arbitrary parameters in order to acquire determinate values for the distribution of thermal and mechanical energy illustrates the sense in which reality can be thought of as no more determinate than fundamental reality. Put simply, no specific amount of mechanical energy is implied by fundamental reality even though all the fundamental attributes are absolutely precisely defined. Instead, there is only a conditional of the form, "For any choice of rest *R* and choice of corpuscle groupings *A* and *B*, there will be a determinate value for the thermal energy and mechanical energy." If there were some brute fact of the matter about the amount of thermal or mechanical energy that went beyond what was already implied by one's theory of fundamental reality plus any fundamentally arbitrary parameters, that would indicate that this brute fact should count as fundamental.

Discussion of principle (4), which claims that fundamental reality is consistent, is deferred until §1.8.

To conclude, let me note several ideas often associated with 'fundamental' that I believe are best kept separate from our general conception of fundamental reality. For one thing, nothing about fundamental reality, as I conceive of it, requires that there are *levels* of reality or degrees of reality beyond the mere distinction between fundamental and derivative. So, many of the criticisms addressed at the idea of a fundamental *level*, like Schaffer (2003), do not apply to my conception of fundamental reality. Nor does my conception of fundamentality require that what is fundamental be small, like a point-like property instance. Nor is it required that what is fundamental be metaphysically simple; a fundamental entity can have complexity and consist of fundamental parts.

Although much more could be said to delineate fundamental from derivative, I hope I have sketched a clear enough distinction in order to be able to make sense of its primary function in my account of causation: to support a certain kind of reductive relationship, which I will now examine.

1.7 Abstreduction

The relation between thermal and mechanical energy and the fundamental attributes of the simple theory of classical mechanics illustrates an important kind of reduction. Unfortunately, 'reduction' is a famously over-used term with so many different interpretations that it cannot be trusted for secure communication. So, in order to minimize the potential for misleading associations with other people's usage, I hereby introduce a proprietary version of the general idea of reduction: ***abstreduction***. A paradigm example of abstreduction is the relation between mechanical (or thermal) energy and the fundamental attributes of the simple theory of classical mechanics. Mechanical energy abstreduces to fundamental reality (under the pretense that fundamental reality answers to the simple theory of classical mechanics).

Reduction is closely associated with reductive explanation. Because explanation is an extremely contentious topic, I want to be clear that although I believe an abstreduction is a legitimate form of reductive explanation, I do not subscribe to any particular theory of explanation nor do I have any ax to grind concerning which explanations count as genuinely reductive. My aim is merely to cite the preceding account of how mechanical and thermal energy are related to the fundamental attributes posited by the simple theory of classical mechanics and then to argue that in whatever sense that account serves as a reductive explanation of mechanical and thermal energy, my metaphysics of causation will incorporate a reductive explanation of the derivative aspects of causation to the fundamental aspects of causation. The only prominent disanalogy is that we know well enough the content of the simple theory of classical mechanics, but at this stage of human history we can only speculate about a correct and complete theory of fundamental physics and thus fundamental reality more generally.

The point of an abstreduction is to abstract away from the details of fundamental reality in a way that allows us to make sense of derivative quantities in terms of fundamental reality and fundamentally arbitrary parameters. It provides a structure for fuzzing fundamental reality. Imagine we start with a particular model of fundamental reality, F, which specifies fundamental laws and specifies how all the fundamental attributes are arranged. Suppose further that we believe in the existence of a certain quantity D whose value is not implied by F. In order to provide an abstreduction of D to F we engage in the following two stage process.

In the first stage, we supplement F with fundamentally arbitrary parameters and provide a function so that the quantity D has a specific value in terms of these parameters together with quantities from F. The illustration of mechanical energy above was meant to demonstrate that E_M can be derived from a choice of rest and a choice of corpuscle groupings in conjunction with the masses and relative speeds present in any model of the simple theory of classical physics.

The mere fact that we can derive some quantity D from a model of fundamental reality F (with any extra parameters we choose) shows nothing interesting by itself. It is trivial, for example, to create some function of the fundamental variables of the simple theory of classical mechanics. We could have invented a quantity, quinergy, defined as $\sqrt{m}v^5$ for any corpuscle, given some standard of rest, and defined for collections of corpuscles by summing their individual quinergies. The reason no one takes quinergy seriously as an existing property, I think, is that it is not a particularly useful scientific quantity. It plays no role in systematizing or explaining the behavior of anything anyone cares about; it is not a conserved quantity; it plays no compelling role in any macroscopic phenomena. What it seems like we need to do in order to justify the status of thermal and mechanical energy as derivative properties is to account for their utility. For instance, we might note that the distinction between mechanical and thermal plays a role in our account of how much energy can be extracted from a system. (For a system at one temperature, only mechanical energy can be extracted.) We might also note that the stability of thermal and mechanical energy over appropriate time scales helps to make them useful.[6]

In the second stage of an abstreduction of D to F, one attempts to explain why the quantity D is a useful magnitude to consider. Unfortunately, it is difficult if not

[6] According to the formula for thermal energy, the thermal energy does vary sharply as the corpuscles slow down when they (nearly) collide with one another, but so long as there are many particles integrated within the same physical system, these brief jiggles in the amount of thermal energy are small relative to the total thermal energy and become negligible if one averages them over suitable time scales.

impossible to formalize how one explains the utility of some quantity. There is no general scheme anyone is aware of for measuring and comparing the usefulness of different quantities. So this stage of the abstreduction is going to involve appealing to common sense and our collective wisdom concerning what quantities are worth positing. This makes the boundary between derivative reality and the non-existent at least as imprecise as our criteria for utility, but I cannot see why this consequence would be problematic. In particular, there is no harm in concluding that quinergy exists (derivatively) but is not worth bothering with. In any case, if we complete both stages, we have completed an abstreduction of D to F.

A brief terminological note is needed here before I discuss how my metaphysics of causation is abstreductive. Because people's prior commitments about causation are so diverse, I prefer to avoid as much as possible referring to a cause-effect relation in my account. To the extent I refer to causation, that is meant non-technically as a way of speaking with the masses. I will instead use the term 'causation-like' for relations that play some of the roles we ordinarily associate with the cause-effect relation. In particular, 'causal relations' misleadingly suggests irreflexivity, asymmetry, and a discrimination between important causes and negligible background factors. It will be important for my account that there are some fundamental relations that can take over the role traditionally played by token cause-effect relations and thus serve as singular causal relations, but it will not be important whether these fundamental causation-like relations are irreflexive, asymmetric, or suitable for distinguishing the relative importance of partial causes.

My main task in this volume is to conduct an abstreduction of general causation to singular causation. I will define some fundamental singular causation-like relations in my account of the bottom conceptual layer and define some derivative general causation-like relations in my account of the middle conceptual layer using some fundamentally arbitrary parameters. Once the metaphysical structures have been defined, it should be obvious how any derivative causation-like relation with a well-defined value gets its unique determinate value from fundamental reality together with the specified fundamentally arbitrary parameters. I then only need to demonstrate the utility of my derivative causation-like relations, which will be accomplished throughout the middle of this volume by appealing to its simplicity, its flexibility, its generality, and its role in the explanation of causal asymmetry.

In order to carry out my abstreduction, I will stick to the following plan. In chapter 2, I will present an account of fundamental causation-like relations. The most important relations in my account of fundamental reality are determination and (a form of) probability-fixing among events. For example, the complete state of the world at one time might determine a later event, or it might determine that some kind of event has a one-third chance of occurring. This form of singular causation is similar to the models of causation proposed by John Stuart Mill with his "real causes" and by J. L. Mackie with his inus account of causation. It also resembles "productive" notions of causation when the laws propagate states continuously through time.

In chapters 3 and 4, I will present an account of derivative causation-like relations, which inhabit the middle conceptual layer of causation. These will include my own variant of a counterfactual conditional and a corresponding notion of counterfactual dependence or difference-making. Traditionally, difference-making accounts of causation have mostly been competitors to determination accounts of causation, but in my theory, the (derivative) difference-making relations are defined in terms of how fundamental laws propagate hypothetical fundamental states through time, whether deterministically or with fundamental chanciness.

Just as I described parameters that allow one to determine the amount of kinetic and thermal and mechanical energy given the totality of corpuscle attributes, I will provide parameters that allow one to determine the magnitude of difference-making (or counterfactual dependence) using any fundamental laws that determine or fix probabilities. This abstreduction allows my metaphysics of causation to quarantine the shiftiness and vagueness of counterfactuals, which have long plagued difference-making accounts of causation.

Summarizing the important points discussed in this section, to abstreduce some quantity D to a model of fundamental reality F involves specifying some fundamentally arbitrary parameters and explaining how these parameters (together with F) make D a determinate quantity with sufficient utility. Abstreduction reveals how an existent D is nothing more than a handy way to abstract away from the details of the fundamental existent F. The goal for my theory of causation is (1) to show how relations of difference-making (or counterfactual dependence) can be defined in terms of fundamental laws and fundamental events using fundamentally arbitrary parameters, and then (2) to show how these relations of difference-making are useful for abstracting away from the fundamental laws governing the detailed motion of matter.

1.8 Strict Standards and Relaxed Standards

In this section, my goal is to clarify how the bottom and middle conceptual layers of causation, which I associate with the metaphysics of causation, differ from the top conceptual layer of causation, which I associate with various non-metaphysical aspects of causation, especially including the role of causation in the special sciences.

The distinction that separates the top layer is a methodological one based on how theoretical inconsistency should be evaluated. While it is widely believed that avoiding contradictions is important for any theory, there are systematic practical differences in how threats to consistency are resolved, differences related to whether a theory concerns fundamental reality. There are several case studies that successfully illustrate how a theory can be inconsistent yet hedged in a way that allows it to provide highly non-trivial predictions as well as acceptable explanations. These include the old quantum theory of black body radiation (Norton 1987), relativistic electromagnetism (Frisch 2005a), and more (Meheus 2002). Implementing a system of managed inconsistency by disallowing a restricted class of troublesome inferences allows us to make sense of theories that are strictly speaking inconsistent or incoherent as complete theories. Despite formal inconsistency when construed as complete, such theories can still succeed in their usual roles of predicting, systematizing, and explaining, by being treated as incomplete or imprecise.

When a theory purports to be a complete fundamental theory and proposes inconsistent rules for its components, we rightly reject the theory out of hand as unacceptable. This practice is justified by virtue of our conception of what it is for a theory to be about fundamental reality. Our theories of fundamental reality forbid contradictions, I think, because of a commitment to the thesis that no matter how inscrutable and paradoxical reality may seem, deep down, there is some consistent way reality is. This is the fourth principle guiding my conception of fundamentality from §1.6, which I expressed as, "Fundamental reality is consistent," by which I meant that fundamental reality obeys some metaphysical correlate to the law of non-contradiction, as discussed for example by Tahko (2009).

When seeking a theory of fundamental physics, we often formulate ***dynamical laws***, which are laws that constrain how the universe evolves. Imagine a fundamental theory that specifies two dynamical laws and that in the special case where there is a corpuscle at rest by itself, the two laws disagree about what will happen. One law dictates that the corpuscle will remain at rest; the other dictates that the corpuscle will oscillate. If the fundamental theory permits the possibility of a corpuscle being at rest, then the theory in effect provides two conflicting rules for what will happen. Such theories are uncontroversially and correctly regarded as unacceptable theories of fundamental reality.

For a more realistic illustration, we can consider the theory of relativistic electromagnetism, whose laws include Maxwell's laws and the Lorentz force law. Maxwell's laws require that electromagnetic charges be treated as a field-like quantity, a sort of charged fluid. The Lorentz force law requires that charges be treated as (discrete) corpuscles. These two requirements are inconsistent, and it is not clear how to tweak them to remove the inconsistency. If the particles are truly point-like, then the electromagnetic field at every particle location is infinitely strong, which disallows the Lorentz force law from defining a finite force on the particle. If the particles are truly field-like, the internal electromagnetic field forces should make the particle explode. One could postulate additional physics to hold the charged fluid bunched together as a particle, but that force would imply the falsity or incompleteness of the laws of electromagnetism as applied to the charge itself. Fortunately, the inconsistency of electromagnetism is adequately *managed* by not using both laws at the same time for the same material and by not demanding that the theory fully address the question of how charged particles self-interact. In this way, the theory can be technically inconsistent as a complete theory of fundamental reality, but also acceptable as an incomplete theory or as a theory of derivative reality that is only approximate and relies on additional resources in fundamental reality to adjudicate what is fundamentally going on.

In order to explore two approaches to the threat of inconsistency, it is useful to introduce some new terminology. Let us say that a theory's rules ***conflict*** when, for some realistic circumstance, they make contradictory attributions. The possibility where a corpuscle must both remain motionless and yet oscillate (relative to the same frame of reference) is a paradigmatic example of a conflict. The meaning of 'realistic circumstance' can vary depending on what kind of theory is being offered. Some theories, like those of fundamental physics, are meant to be rich enough to characterize their own conception of nomological possibility. For such theories I mean to count as realistic any circumstance that is nomologically possible according to the theory itself, regardless of whether it is possible according to the actual laws. (This notion of nomological possibility could take into account restrictions on the kinds of matter allowed and the space-time structure, not just restrictions given by equations of motion or conservation laws.) Other theories, like those in anthropology or food science, do not specify what is nomologically possible but implicitly rely on an imprecise antecedent notion of possibility. For such theories, any situation that is possible according to this antecedent notion counts as a realistic possibility regardless of whether there are any actual laws at odds with the implicit notion. For example, it could turn out that the true laws, whether we know it or not, are so severely restrictive that the only nomologically possible world is the actual world. If so, many possible circumstances entertained by geneticists and economists are not nomologically possible, but our standards for evaluating theories of genetics and economics are such that we treat seemingly realistic circumstances as nomologically possible. The purpose of distinguishing realistic from unrealistic circumstances is to mark the fact that some possibilities are so epistemically remote that we should not care about whether our theory's principles conflict there. For example, we would rightly not reject an otherwise splendid theory of physics merely by virtue of its having principles that conflict in models with a 43-dimensional space-time unless there were some reason to think the actual space-time has 43 dimensions. The same goes for any conflicts a theory might have if we were to countenance the possibility of angels or magic spells. 'Realistic circumstances' is not meant to include all circumstances having an appreciable chance of obtaining. A possibility does not count as unrealistic in my terminology merely because it has an extremely low objective chance; to be unrealistic it must be subjectively very improbable (according to the relevant experts) because of the laws it invokes or the types of matter or space-time it posits.

There is a difference one can draw between rules that have ***apparent conflicts*** and rules that have ***genuine conflicts***. If there are additional principles in a theory that specify how to ameliorate apparent conflicts, then the theory's rules do not genuinely conflict. There are several legitimate ways to ameliorate apparent conflicts to show they are not genuine conflicts. If a theory has two rules that seem to conflict, one could supplement the theory with conditions restricting the circumstances under which each applies so that for any specific circumstance only one of the rules is operative. Alternatively, one could weaken the content of the theory with a qualifying clause so that whenever the rules conflict, neither is operative. Another option is to augment the theory with a qualifying clause so that whenever rules one and two conflict, only rule one is operative. Many of these maneuvers make a theory less appealing, but in general, a theory could have conflicts that only occur in restricted circumstances, and the point of the amelioration clauses would be to establish explicitly what the theory says in every potentially conflicting circumstance so that the theory's apparent conflicts are never genuine. If a theory refuses to clarify what to do in cases where its rules superficially conflict, or it merely claims that there is always some further resolution to the conflict but does not specify the additional structure that resolves the discrepancy, then that theory has a genuine conflict.

Let us now say that any intellectual discipline whose theories are required to avoid genuine conflicts obeys STRICT standards and any intellectual discipline that allows theories to possess genuine conflicts obeys RELAXED standards.

For illustration, imagine a crude psychological theory of our implicit food concept that offers us the following two rules of thumb for when something counts as food.

1. Something is food if and only if it is a substance of the kind humans serve to each other as something to be eaten.
2. Something is food if and only if it is nutritious.

These rules conflict because it is easy to imagine a substance that would be routinely served at meals but which has no nutritional value, or something that is nutritional but which people find objectionable to eat. A theory that tries to provide an account of our ordinary food concept (as part of psychology) is not normally pretending to provide rules that are strict necessary and sufficient conditions. Instead, the necessary and sufficient conditions expressed above are meant to characterize informal heuristics or rules of thumb that link our thoughts about food with other concepts. Their purpose is to make sense of the following kinds of regularities. When people are presented with information that some S is nutritionally harmful, they tend to think of it as not being food; with information that S is nutritionally beneficial, they tend to think of it as food. There is a default expectation in how we interpret such psychological theories that when a test subject is put into a situation where these default heuristics conflict, additional facts can bear on whether the subject identifies S as food. A psychologist who wanted to flesh out such a theory would provide a more thorough account of the factors that affect whether S is categorized as food, including predictions about which circumstances result in people becoming less certain of their judgments. Yet, we know from experience that the quantity of factors needed to provide precise predictions may be far too large for a practical theory. A much more precise and accurate theory of our folk food concept would presumably need to account for cultural backgrounds, personal differences in gustatory abilities, hunger, how accommodating the subject is to others' judgments, and many other factors. To identify with great precision across a wide range of varying conditions whether a given person will consider *S* as food requires many more facts that are rightly considered outside the scope of psychology. There is also undoubtedly a tradeoff between predictive accuracy and the number of parameters such a theory would need to incorporate. For such practical reasons, the kind of psychology that deals with our concepts in a way that is fairly remote from its neurological implementation could and should be understood as proceeding under RELAXED standards that permit theories to conflict on some assessments of realistic circumstances.

The same considerations apply to so-called ceteris paribus laws appearing in the special sciences. A theory of ecological genetics, for example, might postulate a law that when new islands are formed near populations, the number of species will increase, and another law that when a cataclysm occurs, the number of species will decrease. These laws conflict because there could be a cataclysmic flurry of volcanic eruptions that create new islands. This superficial conflict—that the number of species would both increase and decrease—does not warrant the rejection of a theory that posits both laws. Such laws are not intended as inviolable dictates of nature but as useful rules of thumb that can be overridden in some circumstances. It is also understood that whether the number of species goes up or down depends on the nature and severity of the volcanic activity, the number of islands created, and many other factors that ecological genetics is simply not in the business of accounting for in detail. Because ecological geneticists are not obligated to spell out all the parameters that would ameliorate all apparent conflicts in its models, we can say that ecological genetics obeys RELAXED standards.

When a theory concerns fundamental reality, it is appropriate to hold it to STRICT standards. In most (but not all) cases, when a theory concerns only derivative reality, it is arguably appropriate to hold it only to RELAXED standards. For an example of a case where the subject matter uncontroversially concerns derivative reality but ought to obey STRICT standards, consider the narrow subset of thermodynamics that deals with the distinction between mechanical and thermal energy, again under the pretense that fundamental reality is completely and correctly described by the simple theory of classical mechanics. I know of no reason to think that thermodynamics generally has to hold to STRICT standards, but in the special case of theories that deal only with concepts that have a close enough fit with fundamental reality, it is reasonable to maintain STRICT standards. If we have accepted that the mechanical and thermal energy of macroscopic objects are so closely related to the fundamental attributes of classical mechanics that it does not require the services of any other scientific discipline, we are justified in demanding that a theory rule out any genuine conflicts in its pronouncements regarding thermal and mechanical energy. For example, if a theory claims that very nearly two percent of the energy of any large body of water is thermal, and another part of the same theory claims that very nearly thirty percent of the energy of salt water is thermal, it is appropriate to demand some account of how these two claims are compatible when applied to an ocean, which is both large and salty. An outcome that is uncontroversially unacceptable is for the theory to declare that both percentages are accurate general rules of thumb while remaining silent on what the theory claims about the ocean's thermal energy.

What makes this subset of thermodynamics different from the case of ecological genetics is that the predictions of ecological genetics depend on factors well outside the scope of ecological genetics. We know that conflicts between the law that islands increase the number of species and the law that cataclysms decrease the number of species can be ameliorated by considering a richer set of facts addressed by physics that can ultimately settle whether the number of species increases or decreases in any particular case and over what time scales. But for the physicist whose theory refers to thermal energy, there is no further discipline to which conflicts can be delegated. If we have accepted that classical mechanics is our fundamental theory and that thermal energy is abstreduceable to the fundamental attributes of the corpuscles, the only resources available to ameliorate the apparent conflict are the additional parameters that make thermal energy determinate. The difference between STRICT and RELAXED standards is just that any theory adhering to STRICT standards cannot just hand-wavingly assert that there are further details that ameliorate apparent conflicts. It must explicitly state the parameters that ensure its concepts are being applied consistently. The reason for holding STRICT standards in this special subset of thermodynamics is that we already hold STRICT standards for theories of fundamental reality and we have committed ourselves to the thesis that there are no further facts that some other discipline could supply that would ameliorate the conflict. If we were to abandon belief that there is a fundamental physics or commit ourselves to a different fundamental physics that makes the relation between it and energy more opaque, that could motivate us to adopt RELAXED standards for this portion of thermodynamics.

Throughout the preceding discussion, I have not in any way ruled out the possibility that some special sciences ought to operate under STRICT standards. I am only providing a few examples where I think it is intuitively plausible that the appropriate standards to hold are RELAXED. The kind of psychological theory that tries to model our implicit beliefs about some concept like 'food' or 'causation' is operating at a fairly high level of abstraction, high enough so that its pronouncements can in principle be perfectly acceptable qua high-level theory of our concepts even though it does not provide parameters that guarantee a lack of conflict. This is entirely compatible with the possibility that some other kinds of psychology need to obey STRICT standards. Also, I do not intend my terminology to insinuate that theories operating under RELAXED standards are in any way less respectable than theories that are STRICT or that there is any less rigor in disciplines that employ RELAXED standards. The distinction between STRICT and RELAXED is merely a device to help delegate responsibility among disciplines for ameliorating conflicts.

Although it is difficult to rigorously defend hypotheses about which disciplines ought to hold STRICT rather than RELAXED standards, I do think it is a fair characterization of the intellectual activity known as metaphysics that people engaging in it believe metaphysical theories should obey STRICT standards, and I think they are correct to do so. Although 'metaphysics' is a term with evolving and contentious meanings, metaphysics is uncontroversially the general study of reality. In particular, theories of metaphysics are aimed at an account of reality that is not merely a patchwork of conflicting rules of thumb but a more systematic structure that is ultimately consistent. The motivation for adhering to STRICT standards in metaphysics makes sense given that foundational role of metaphysics does not permit it to delegate conflicts to other disciplines. Conflicts within theories of the special sciences often do not need explicit amelioration because there are virtually always additional physical facts not subsumed by the special science in question that one can plausibly appeal to for amelioration, but metaphysics has no other discipline available to ameliorate its conflicts. Metaphysics does often delegate to other disciplines to fill in some details. For example, a metaphysical theory might pronounce on what kinds of properties are possible and then task biology with discovering which particular biological properties exist. But it is not the role of biology to clarify the conditions under which the metaphysical theory's characterization of properties would be inapplicable or overridden by some alternative. A special science might reveal inadequacies of a metaphysical theory of properties, but it wouldn't provide a richer story about the general nature of properties than what metaphysics itself is expected to provide. In that sense, it is appropriate to hold metaphysical theories accountable to STRICT standards.

The point of this section has been to introduce some theoretical machinery so that I can now state a conclusion that I will eventually defend in chapter 8. Although any theory concerning the metaphysics of causation should obey STRICT standards, there is an activity commonly regarded by philosophers as part of the metaphysics of causation that can be entirely adequate even if it only satisfies the weaker RELAXED standards. This activity is the provision of rules for when a singular event counts as "one of the causes" of some chosen event. Sometimes, such rules are known as theories of ***actual causation***, though I forego this terminology because of its incorrect implication that less noteworthy singular causes are not part of the actual world. Weakening the conditions of adequacy for theories governing such singular causes makes it easier to understand the range of intuitions that have long been considered by philosophers as touchstones for identifying the correct metaphysics of causation. What my account does, in effect, is to relocate some aspects of causation that have traditionally been understood as metaphysical, like preëmption, to the non-metaphysical aspects of causation. The purpose of the distinction between STRICT and RELAXED is to mark the boundary between those aspects of causation that need to be systematized in a principled and fully consistent system and those that do not. Chapter 9 will illustrate how our intuitions about singular causation could be systematized in a principled and explanatory theory with genuine conflicts.

1.9 Limitations on the Aspirations of Empirical Analysis

Putting together the new terminology from the previous two sections, I can now complete my presentation of empirical analysis by pointing out an important restriction on what activity is needed to produce an adequate empirical analysis. In §1.1, I initially characterized the goal of an empirical analysis of *X* as identifying "scientifically improved concepts of *X*." However, when we are engaged in an empirical analysis that concerns an *X* that is rightly considered part of *metaphysics*, as elucidated above, the scope of the project is automatically limited in the sense that one is not required to provide a regimentation to serve as an improvement for *all* appearances of *X* in the sciences. In particular, my empirical analysis of the metaphysics of causation is not required and is not intended to address all the locations where causal terminology is invoked. Quite to the contrary, it is intended to accomplish the much narrower task of connecting *all* causal regularities in the special sciences to fundamental reality (simplified provisionally as fundamental physics) in a STRICT manner. Some readers might have thought on the basis of my earlier discussion that an empirical analysis of causation requires an examination of the many uses of causal terminology or the variety of causal principles invoked in the special sciences, but such thoughts are incorrect. An empirical analysis of the metaphysics of causation only needs to develop concepts needed for the STRICT connection between derivative reality and fundamental reality. The omitted discussion is a task for an empirical analysis of the non-metaphysical aspects of causation.

It ought to go without saying that this methodological division of labor imposed by empirical analysis does not in any way denigrate special sciences or cast doubt on the importance of the full range of causal principles and causal concepts used in the special sciences. Rather, the consequence of this maneuver is in general to insulate the practices of the special sciences from the details of fundamental reality and in particular to grant them wide latitude to use a variety of causal concepts without having to draw any explicit connection to fundamental physics. This is analogous to the division of labor induced by the portion of physics that abstreduces a limited class of energy types, like thermal and mechanical, to fundamental attributes. By making explicit how every invocation of energy within a limited class of energy types can in principle be connected to fundamental reality in a consistent way, other special sciences are thereby freed to mention and build on these energy types without having to provide a maximally detailed account of what these energies consist of.

An ecologist, for example, might want to discuss energy flow through trophic levels by referring to the amount of energy held in plants that is available for appropriation by herbivores. Using the division of labor marked by the STRICT and RELAXED distinction, the ecologist would need to ensure that her energy concepts are well enough managed so that ecological processes she models do not violate conservation of energy or permit perpetual motion machines. But, crucially, she would not be required to be maximally precise in her account of what the boundary is between, say, plant energy and bacterium energy. That is a welcome consequence, given that plants are laden with bacteria. The RELAXED standards enforce enough linkage between ecology and fundamental physics to ensure that ecology does not violate laws of physics but otherwise leaves ecology free to posit forms of energy without having to express them as an explicit function of the attributes of fundamental physics. In general, RELAXED standards permit managed inconsistency.

Similarly, the metaphysics of causation I will provide will not be of any practical use to researchers who seek the causes of cancer. Nor will it directly address how to conduct causal modeling projects. Its purpose instead is to serve as a universal basis to which the special sciences can defer in order to backstop their use of not-maximally-precise causal terminology. With my account in place, the ecologist will be free to attribute the decline of tiger populations to the human appropriation of its prey without such claims hinging on the contentious question of whether fundamental reality includes something more than physics or the contentious question of whether causation is ontologically more than matter evolving according to laws of physics. I fully recognize that such worries are not pressing to practicing scientists, but it is of central concern to the long-standing philosophical question of whether and how there could be a relatively sparse model of fundamental physics that is sufficient for all of reality.

1.10 Comparison of Empirical and Orthodox Analysis

In order to illustrate the practical import of the distinction between STRICT and RELAXED, I will now emphasize how my pair of empirical analyses differ from orthodox analyses in their approach to singular causation.

Although orthodox analyses of causation have varying overall goals, one of the recognized tasks for any orthodox analysis is to identify non-trivial rules for which events count as causes, given not-too-causally-loaded information about the laws of nature and the history of occurrent facts. When we cite instances of causation—a whale breach causing a splash, for example—we intend to draw special attention to a small portion of the universe as being important to the effect. These events are called "the causes" of the effect or, more recently, the "actual causes" of the effect.

Orthodox theorizing about causation is expected to provide rules for what makes something count as one of these causes.

The singular causes sought by orthodox analyses are typically not fantastically detailed physical states but are intended to be the kinds of events people cite when asked about the causes of some particular event. For example, they might mention the launching of a ship, the loss of one tooth, or an increase in the gross domestic product during the fourth quarter of 1968. From here on, I will refer to such events as ***mundane events***. Orthodox accounts of singular causation focus on relations among mundane events even when they allow that causal relations can exist among events that are physically sophisticated like the complete microphysical state existing on an infinitely extended time slice. Because these sophisticated kinds of events might play a role in singular causation, it is valuable to distinguish the kind of singular causes that are typically mundane events. Let us say that a ***culpable cause*** of some event *e* is an event that counts as "one of the causes of *e*" in the sense employed by metaphysicians who study causation. 'Culpable cause' is not a technical term but merely a label for the "egalitarian" (Hall 2004) notion of cause that orthodox metaphysicians seek when they ask, "What are the causes of (the singular event) *e*?" I emphasize that 'culpable cause' is my proprietary expression[7] introduced to reduce confusion about what 'cause' by itself connotes. Two further qualifications can be made at this point. First, culpable causes are so named because they are events that are blameworthy for the effect, but the terminology is not meant to imply that our intuitions about the relevant notion of singular cause absolutely perfectly matches our intuitions about how to attribute causal blame. Second, there is perhaps an ambiguity in the expression "a cause of *e*." It could mean "one of the causes of *e*" or it could mean "something that caused *e*." These are not always recognized as equivalent. When Guy won the lottery, his purchase of the ticket was one of the causes of his winning, but people would not normally say that purchasing the ticket caused Guy to win. 'Culpable cause' refers to the 'one of the causes' disambiguation.[8]

One feature that makes the orthodox analysis of causation a project in metaphysics rather than armchair psychology is that a proper analysis is required to provide a principled account of what is common to all cases of causation. Imagine that a psychologist offers a theory of causation consisting of a list of eight exemplars of the cause-effect relation and thirteen exemplars of the lack of a cause-effect relation. The theory says *c* is a cause of *e* if and only if the situation where *c* and *e* happen is closer to one of the positive exemplars than to any of the negative exemplars, closeness being judged by one's own intuitive off-the-cuff assessment of similarity. A theory of this form might make for an interesting psychological

[7] The term 'culpable cause' has been used previously by Mark Alicke (1992) to designate something altogether different: the psychological effect of perceived moral blameworthiness on judgments of causal impact.

[8] Some professionals report detecting no ambiguity here. I am fairly certain that if 'a cause' does not strike you as ambiguous, it is already picking out the intended conception of culpable cause.

theory and might even accrue empirical support if our causal reasoning is based less on rules than on some sort of pattern-matching capacity. But from the perspective of metaphysics, it would fail to capture what is similar in all cases of causation in an appropriate way. Such theories are merely fitting the data, whereas the metaphysician is interested in a theory based on principles that would connect causation with laws, chance, time, and would more closely resemble necessary and sufficient conditions.

Another feature that distinguishes an orthodox analysis of causation is the expectation that it is to be held to STRICT standards of consistency. For illustration, consider the following crude theory of our concept of causation, which is meant to parallel the crude theory of our concept of food from §1.8.

1. An event *c* is a cause of *e* iff *c* raises the probability of *e*.
2. An event *c* is a cause of *e* iff there exists a chain of probability-raising relations going from *c* to *e*.

This conjunction of rules might be faulty for multiple reasons, but let us focus just on realistic possibilities where the rules conflict. In an example (Suppes 1970) attributed to Deborah Rosen, a golfer's slice, *c*, lowered the probability of a good shot, *e*, and so was not a cause of *e* according to the first rule. But the slice did raise the probability of hitting a tree, which in turn raised the probability of the ball bouncing back in a better position, thus making *c* a cause of *e* according to the second rule.

By the RELAXED standards appropriate to most special sciences, including the kind of psychology concerned with modeling people's responses to questions about what caused what, it is acceptable for a theory to claim that people employ both biconditionals as rough-and-ready heuristics for assessing the existence of a cause-effect relation. Under RELAXED standards, having multiple conflicting rules for what events count as causes can be acceptable even if there is no further account in the theory of how to resolve (for all realistic circumstances) which heuristic is operative.

From a metaphysical point of view, such conflicting rules are unsatisfactory as an account of causation. In metaphysics, one is thinking of the causes as some element of external reality. A theory that provides conflicting pronouncements about whether *c* is a cause and provides no further resources to settle which rule is applicable and fails to relativize the incompatible facts to parameters that would remove the conflict, is in effect stating that its model of the actual world is inconsistent, which is uncontroversially unacceptable. One of the crucial standards by which orthodox metaphysical theories are to be judged is that their rules for causation need to be consistent. Furthermore, one is not allowed to save the inconsistent rules merely by adding a hand-waving qualifier that says, "In some cases, the first rule holds, and in other cases, the second rules holds." For metaphysical theories, room is typically permitted for vagueness by allowing a theory to avoid issuing a determinate judgment in all cases, but there is an obligation to ensure the theory does not judge that in a single scenario, *c* is both a cause of *e* and not a cause of *e*.

Orthodox accounts of causation attempt to find rules for attributing causation that on the one hand are principled and obey STRICT standards of adequacy, and on the other hand closely fit the relevant psychological data, including informed judgments about which partial causes should count as explaining the effect. What my account does is to replace this project with two empirical analyses. The empirical analysis of the metaphysics of causation is supposed to be principled and STRICT but is not supposed to fit any psychological data in the sense of rendering people's judgments about culpable causation explicitly true. The empirical analysis of the non-metaphysical aspects of causation is intended to be principled and fit the psychological data in the sense of systematizing common judgments about culpable causation, but it only needs to satisfy RELAXED standards to count as adequate for its intended purpose. This pair of empirical analyses accomplishes what the orthodox approach attempts to do in a unified treatment, but because it segregates the needed concepts into a metaphysical part and a non-metaphysical part, it is able to optimize metaphysical concepts in accord with the demands of fundamental reality and non-metaphysical concepts in accord with the demands of folk psychology or epistemology or whatever practices in the special sciences one wishes to consider. It is thereby able to achieve greater optimization without significant loss.

1.11 Summary

To conclude this introductory chapter, I will now return attention to the three conceptual layers of causation described in §1.4 and summarize how each layer of my tripartite account of causation is intended to work. Each of the three conceptual layers of causation, depicted in Table 1.2, contains its own causation-like concepts, none of which needs to match what we pre-theoretically think of as causation. Yet, all three layers together allow us to make sense of everything concerning causation that we need to make sense of.

The bottom layer contains concepts that are not tailored to match our everyday causal talk but are supposed to provide just enough structure to support the work

TABLE 1.2 The three conceptual layers of causation.

Layer	Subject	Metaphysical status	Standards of adequacy
Top	Non-metaphysical aspects	Derivative	RELAXED
Middle	Derivative metaphysics	Derivative	STRICT
Bottom	Fundamental metaphysics	Fundamental	STRICT

of the middle and top layers in explaining the utility of folk causal talk and vindicating causal principles in the special sciences. It does so mainly by guaranteeing the existence of a consistent basis of facts to which the special sciences can delegate the amelioration of conflicts, thus freeing the special sciences to employ concepts that are not completely, explicitly, consistently systematized and connected to fundamental reality.

The fundamental causation-like relations serve as the foundation for all causal relations and are as objective as causation ever gets. In chapter 2, I will lay out the details of these fundamental causation-like relations, which are based on the hypothesis that some chunks of fundamental reality—events that instantiate fundamental particles or fields of some sort—fix objective probabilities for (or determine) other chunks of fundamental reality. These relations of probability-fixing and determination serve as singular causal relations in my metaphysics of causation.

The middle layer abstracts away from the fundamental relations by incorporating parameters that fuzz the fundamental details in order to represent the sort of behavior we humans deal with in the special sciences and in everyday life. This helps to account for why some kinds of events reliably bring about characteristic effects. The generality that smoking causes cancer, for example, will be understood on my account in terms of derivative metaphysical relations. Most relevant smoking events fix a higher probability of acquiring cancer than the probability fixed by relevant non-smoking events. Relations of probability-raising and related forms of probabilistic influence exist only relative to parameters that are fundamentally arbitrary. These parameters characterize how to fuzz the microphysics and specify counterfactual scenarios to help contrast how things actually evolve with how things could have evolved. Because there is no unique correct way to set the values of these parameters, the relations that incorporate them do not correspond to fundamental reality but involve some degree of arbitrariness, just like mechanical and thermal energy. Nevertheless, the difference-making relations holding in the middle layer are not independent of the bottom layer. The determination and probability-fixing relations from the bottom layer constitute the basic materials for quantifying difference-making, which acquires determinate values once the designated fundamentally arbitrary parameters are assigned values.

In the end, my metaphysics of causation incorporates several common themes in the causation literature: difference-making, nomic dependence, and production.

Together, the middle and bottom layers support a scientific account of the empirical phenomena associated with effective strategies. How they do so is the subject of chapter 5. I will follow the general explanation of effective strategies with a proof of causal directness in chapter 6, and then an account of causal asymmetry in chapter 7. These chapters will exploit the technical terminology developed in chapter 2 to demonstrate important characteristics of causation that hold by virtue of fundamental laws of physics and thus bolster the hypothesis that causation is at least partly based on fundamental physics.

The purpose of the top layer is to provide an account of those aspects of our causal concepts that are inessential to the explanation of the metaphysics of causation but are important components of causation and causal explanation. The main concept present in this layer is the notion of a culpable cause expressed in statements like "The intruder caused the dog to bark," and "Oppressive heat was one of the causes of the traffic jam." I believe the primary (though not sole) reason we have this kind of causal concept is that it allows us to grasp the important metaphysical relations in a cognitively convenient form. Ideally, we could figure out what kinds of strategies are effective by running lots of controlled studies with a large sample of initial conditions that are tailored to expose how much difference each aspect of reality makes in bringing about any desired effects. But because humans need to gain knowledge of effective strategies even when such studies are not feasible and because our ancestors needed causal knowledge when they did not know how to run controlled studies, we have evolved cognitive shortcuts that allow us to make good guesses about general causal relations from an impoverished data set. This, I contend, is one good reason for our having strong intuitions concerning relations of singular causation. Our concept of culpable causation on the whole does a good enough job of tracking the causation-like relations from the bottom and middle layers for practical purposes, but because the folk conception of causation incorporates additional epistemic features that play no essential role in accounting for effective strategies, there are some significant mismatches. Our instinctive judgments of causation will often enough identify *c* as a non-cause of *e* when *c* is generally useful for bringing about events of the same chosen type as *e*. Yet, the main reason we have a concept of causation is that we need to distinguish between the kinds of events that are generally effective at achieving desired results. In chapter 8, I will attempt to explain why this folk causal notion is unneeded in an account of the metaphysics of causation, and then in chapter 9, I will provide a simplistic psychological theory in order to demonstrate why it makes sense for us to have this folk conception of causation given that the metaphysics of causation obeys the system I lay out in the rest of the book.

In the end, the conceptual system I advocate proves to be an extremely revisionary model of our ordinary understanding of causation. It involves reconfiguring the relation between singular and general causation, abandoning traditional models of counterfactual dependence, modifying the accepted distinction between causal and non-causal statistical correlations, and even dispensing with the dogma that we are unable to influence the past. A radical architecture for causation is hardly surprising, though, since the account focuses on explaining empirical phenomena and refuses to be held captive to common sense.

{ REFERENCES }

Adams, E. (1975). *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: D. Reidel.

Adams, E. (1976). "Prior Probabilities and Counterfactual Conditionals," in W. L. Harper and C. A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* 1, 1–21.

Albert, D. (2000). *Time and Chance*. Cambridge: Harvard University Press.

Alicke, M. (1992). "Culpable Causation," *Journal of Personality and Social Psychology* 63 (3), 368–378.

Allori, V., Goldstein, S., Tumulka, R., Zanghi, N. (2008). "On the Common Structure of Bohmian Mechanics and the Ghirardi-Rimini-Weber Theory," *The British Journal for the Philosophy of Science* 59, 353–389.

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

Austin, J. L. (1961). "A Plea for Excuses," in *Philosophical Papers*, 123–152. Oxford: Oxford University Press.

Bechtel, W. and Richardson, R. (1993). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. Princeton, NJ: Princeton University Press.

Beebee, H., Hitchcock, C., and Menzies, P., eds. (2009). *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Beebee, H. (2004). "Causing and Nothingness," in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Beisbart, C. and Hartmann, S., eds. (2011). *Probabilities in Physics*. Oxford: Oxford University Press.

Bell, J. (1981). "Bertlmann's Socks and the Nature of Reality," *Journal de Physique*, Colloque C2, suppl. au numero 3, Tome 42, C2 41–61; reprinted in Bell, J. (2004). *Speakable and Unspeakable in Quantum Mechanics*, 2nd ed. Cambridge: Cambridge University Press, 139–158.

Benton, M. (2005). *When Life Nearly Died: The Greatest Mass Extinction of All Time*. New York: Thames and Hudson.

Block, N. and Stalnaker, R. (1999). "Conceptual Analysis, Dualism, and the Explanatory Gap," *The Philosophical Review* 108 (1), 1–46.

Boltzmann, L. (1895). *Nature* 51, 413.

Brewer, R. and Hahn, E. (1984). "Atomic Memory," *Scientific American* 251 (6), 36–50.

Campbell, J. K., O'Rourke, M., and Silverstein, H., eds. (2007). *Causation and Explanation*. Cambridge, MA: MIT Press.

Carroll, J. (1991). "Property-level Causation?" *Philosophical Studies* 63, 245–270.

Carroll, J. (1994). *Laws of Nature*. Cambridge: Cambridge University Press.

Carroll, J. (2004). *Readings on Laws of Nature*. Pittsburgh: University of Pittsburgh Press.

Cartwright, N. (1979). "Causal Laws and Effective Strategies," *Noûs* 13, 419–437. Reprinted in N. Cartwright, *How the Laws of Physics Lie*. Oxford: Clarendon Press, 1983, 21–43.

Cartwright, N. (1994). *Nature's Capacities and their Measurement* Oxford: Oxford University Press.

Chalmers, D. and Jackson, F. (2001). "Conceptual Analysis and Reductive Explanation," *The Philosophical Review* 110 (3), 315–360.

Collingwood, R. G. (1940). *An Essay on Metaphysics*. London: Oxford University Press.

Collins, J., Hall, N., and Paul, L. A., eds. (2004). *Causation and Counterfactuals*. Cambridge: MIT Press.

Dainton, B. (2001). *Time and Space*. Montreal: McGill Queen's University Press.

Diaconis, P. and Engel, E. (1986). "Some Statistical Applications of Poisson's Work," *Statistical Science* 1 (2), 171–174.

Dowe, P. and Noordhof, P. (2004), eds. *Cause and Chance: Causation in an Indeterministic World*. London: Routledge.

Dowe, P. (1992a). "An Empiricist Defence of the Causal Account of Explanation," *International Studies in the Philosophy of Science* 6, 123–128.

Dowe, P. (1992b). "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory," *Philosophy of Science* 59, 195–216.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Dowe, P. (2009). "Causal Process Theories," in H. Beebee, C. Hitchcock, and P. Menzies (eds.), *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Dretske, F. (1977). "Laws of Nature," *Philosophy of Science* 44 (2), 248–268.

Driver, J. (2008a). "Attributions of Causation and Moral Responsibility," in W. Sinnott-Armstrong (Ed.) *Moral Psychology (Vol. 2): The Cognitive Science of Morality: Intuition and Diversity* 423–439. Cambridge, MA: MIT Press.

Driver, J. (2008b). "Kinds of Norms and Legal Causation: Reply to Knobe and Fraser and Deigh," in W. Sinnott-Armstrong (ed.), *Moral Psychology (Vol. 2): The Cognitive Science of Morality: Intuition and Diversity* 459–461). Cambridge, MA: MIT Press.

Ducasse, C. J. (1926). "On the Nature and the Observability of the Causal Relation," *The Journal of Philosophy* 23 (3), 57–68.

Dummett, M. (1964). "Bringing About the Past," *The Philosophical Review* 73 (3), 338–359.

Earman, J. (1986). *A Primer on Determinism*. Dordrecht: Reidel.

Earman, J. (1995). *Bangs, Whimpers, Crunches, and Shrieks*. Oxford: Oxford University Press.

Edgington, D. (2004). "Counterfactuals and the Benefit of Hindsight," in P. Dowe and P. Noordhof (eds.), *Cause and Chance: Causation in an Indeterministic World*. London: Routledge.

Eells, E. (1991). *Probabilistic Causality*. Cambridge: Cambridge University Press.

Elga, A. (2007). "Isolation and Folk Physics," in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press.

Fair, D. (1979). "Causation and the Flow of Energy," *Erkenntnis* 14, 219–250.

Ernst, G. and Hüttemann, A., eds. (2010). *Time, Chance, and Reduction: Philosophical Aspects of Statistical Mechanics*. Cambridge: Cambridge University Press.

Field, H. (2003). "Causation in a Physical World," in M. Loux and D. Zimmerman (eds.), *The Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 435–460.

Fisher, R. (1959). *Smoking: The Cancer Controversy*. London: Oliver & Boyd.

Friedman, M. (1983). *Foundations of Space-Time Theories: Relativistic Theories and Philosophy of Science*. Princeton: Princeton University Press.

Frigg, R. (2009). "Typicality and the Approach to Equilibrium in Boltzmannian Statistical Mechanics," *Philosophy of Science* 76, Supplement, S997–1008.

Frigg, R. (2011). "Why Typicality Does Not Explain the Approach to Equilibrium," in M. Suarez (ed.), *Probabilities, Causes and Propensities in Physics*, Synthese Library, Vol. 347. Berlin: Springer, 77–93.

Frisch, M. (2005a). *Inconsistency, Asymmetry, and Non-Locality*. New York: Oxford University Press.

Frisch, M. (2005b). "Counterfactuals and the Past Hypothesis," *Philosophy of Science* 72, 739–750.

Frisch, M. (2007). "Causation, Counterfactuals and Entropy," in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press.

Frisch, M. (2010). "Does a Low-Entropy Constraint Prevent Us from Influencing the Past?" in G. Ernst and A. Hüttemann (eds.), *Time, Chance, and Reduction*. Cambridge: Cambridge University Press.

Galilei, G. (1960). *De Motu* (I.E. Drabkin, Trans.). Madison: The University of Wisconsin Press. (Original work written 1590)

Ghirardi, G. C., Rimini, A., and Weber, T. (1986). "Unified Dynamics for Microscopic and Macroscopic Systems," *Physical Review D* 34, 470–491.

Glennan, S. (1996). "Mechanisms and the Nature of Causation," *Erkenntnis* 44 (1), 49–71.

Glennan, S. (2002). "Rethinking Mechanistic Explanation," *Philosophy of Science* 69 (3): S342–S353.

Glennan, S. (2009). "Mechanisms," in H. Beebee, C. Hitchcock, and P. Menzies (eds.), *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Glymour, C. and Wimberly, F. (2007). "Actual Causes and Thought Experiments," in J. K. Campbell, M. O'Rourke, and H. Silverstein (eds.), *Causation and Explanation*. Cambridge, MA: MIT Press.

Godfrey-Smith, P. (2012). "Metaphysics and the Philosophical Imagination," *Philosophical Studies* 160 (1), 97–113.

Good, I. J. (1961). "A Causal Calculus I," *The British Journal for the Philosophy of Science* 11, 305–18.

Good, I. J. (1962). "A Causal Calculus II," *The British Journal for the Philosophy of Science* 12, 43–51.

Gold, T. (1962). "The Arrow of Time," *American Journal of Physics* 30, 403–10.

Goodman, N. (1947). "The Problem of Counterfactual Conditionals," *The Journal of Philosophy* 44, 113–28.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Goldstein, S. (2010). "Boltzmann's Approach to Statistical Mechanics," in J. Bricmont, D. Dürr, M.C. Gallavotti, G. Ghirardi, F. Petruccione, and N. Zanghì (eds.), *Chance in Physics: Foundations and Perspectives*, Springer, New York, 39–54.

Goldstein, S. "Bohmian Mechanics", The Stanford Encyclopedia of Philosophy (Spring 2013 Edition), Edward N. Zalta (ed.), URL = ⟨http://plato.stanford.edu/archives/spr2013/entries/qm-bohm/⟩.

Hall, N. (2000). "Causation and the Price of Transitivity," *Journal of Philosophy* 97, 198–222.

Hall, N. (2004). "Two Concepts of Causation," in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*, 2004. Cambridge: MIT Press.

Hall, N. (2007). "Structural Equations and Causation," *Philosophical Studies* 132, 109–136.

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

Hesslow, G. (1981). "Causality and Determinism," *Philosophy of Science* 48, 591–605.

Hitchcock, C. (1993). "A Generalized Probabilistic Theory of Causal Relevance," *Synthese* 97, 335–364.

Hitchcock, C. (1995). "Discussion: Salmon on Explanatory Relevance," *Philosophy of Science* 62, 304–320.

Hitchcock, C. (1996a). "Farewell to Binary Causation," *Canadian Journal of Philosophy* 26, 267–282.

Hitchcock, C. (1996b). "The Role of Contrast in Causal and Explanatory Claims," *Synthese* 107 (3), 395–419.

Hitchcock, C. (2001). "The Intransitivity of Causation Revealed in Equations and Graphs." *Journal of Philosophy* 98, 273–299.

Hitchcock, C. (2003). "Of Humean Bondage," *The British Journal for the Philosophy of Science* 54, 1–25.

Hitchcock, C. (2004). "Do All and Only Causes Raise the Probabilities of Effects?" in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Hitchcock, C. (2007). "Three Concepts of Causation," *Philosophy Compass* 2/3: 508–516.

Hitchcock, C. (2009). "Structural Equations and Causation: Six Counterexamples," *Philosophical Studies* 144, 391–401.

Horwich, P. (1987). *Asymmetries in Time*, Cambridge, MA: MIT Press.

Jackson, F. (1979). "On Assertion and Indicative Conditionals," *The Philosophical Review* 88, 565–89.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Keller, J. (1986). "The Probability of Heads," *The American Mathematical Monthly* 93 (3), 191–197.

Kim, J. (2001). "Physical Process Theories and Token-Probabilistic Causation," *Erkenntnis* 48, 1–24.

Kistler M. (1999). *Causalité et lois de la nature*. Paris: Vrin.

Knobe, J. and Fraser, B. (2008). "Causal Judgment and Moral Judgment: Two Experiments," in W. Sinnott-Armstrong (ed.), *Moral Psychology (Vol. 2): The Cognitive Science of Morality: Intuition and Diversity* 441–447. Cambridge, MA: MIT Press.

Kutach, D. (2001). *Entropy and Counterfactual Asymmetry*, PhD. Dissertation, Rutgers.

Kutach, D. (2002). "The Entropy Theory of Counterfactuals," *Philosophy of Science* 69 (1), 82–104.

Kutach, D. (2007). "The Physical Foundations of Causation," in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press.

Kutach, D. (2010). "Empirical Analyses of Causation," in A. Hazlett, (ed.) *New Waves in Metaphysics*. Palgrave Macmillan, UK.

Kutach, D. (2011a). "Backtracking Influence," *International Studies in the Philosophy of Science* 25 (1), 55–71.

Kutach, D. (2011b). "Reductive Identities: An Empirical Fundamentalist Approach," *Philosophia Naturalis* 47–48, 67–101.

Kutach, D. (2011c). "The Asymmetry of Influence," in C. Callender (ed.), *The Oxford Handbook of Philosophy of Time*. Oxford: Oxford University Press.

Kutach, D. (Forthcoming). "The Empirical Content of the Epistemic Asymmetry," forthcoming in B. Loewer, B. Weslake, and E. Winsberg (eds.), *On Time and Chance*. Cambridge, MA: Harvard University Press.

Kvart, I. (2004). "Probabilistic Cause, Edge Conditions, Late Preemption and Discrete Cases," in P. Dowe and P. Noordhof (eds.), *Cause and Chance: Causation in an Indeterministic World*. London: Routledge.

Lange, M. (2002). *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass*. Oxford: Blackwell Publishers.

Laraudogoitia, P. (1996). "A Beautiful Supertask," *Mind* **105**, 81–83.

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

Lewis, D. (1973b). "Causation," *The Journal of Philosophy* **70** 556–67, reprinted in *Philosophical Papers, Volume 2*, Oxford: Oxford University Press, 1986.

Lewis, D. (1986). "Postscripts to 'Causation,'" *Philosophical Papers, Vol. II* Oxford: Oxford University Press.

Lewis, D. (2000). "Causation as Influence," *Journal of Philosophy* **97** 182–97, reprinted in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Lewis, P. (2006). "GRW: A Case Study in Quantum Ontology," *Philosophy Compass* **1** (2), 224–244.

Loewer, B. (2004). "Humean Supervenience," in J. Carroll, ed. *Readings on Laws of Nature*.

Loewer, B. (2007). "Counterfactuals and the Second Law," in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press.

Loewer, B. (2012). "Two Accounts of Laws and Time," *Philosophical Studies* **160** (1), 115–137.

Machamer, P. and Wolters, G. (2007). *Thinking about Causes: From Greek Philosophy to Modern Physics*. Pittsburgh: University of Pittsburgh Press.

Machamer, P., Darden, L., and Craver, C. (2000). "Thinking about Mechanisms," *Philosophy of Science* **67** (1), 1–25.

Mackie, J. L. (1973). *The Cement of the Universe*. Oxford: Oxford University Press.

Maslen, C. (2004). "Causes, Contrast, and the Nontransitivity of Causation," in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Mather, J. and McGehee, R. (1975). "Solutions of the Collinear Four-Body Problem Which Become Unbounded in a Finite Time," in J. Moser (ed.), *Dynamical Systems, Theory and Applications*, New York: Springer-Verlag.

Maudlin, T. (2004). "Causation, Counterfactuals, and the Third Factor," in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press. Reprinted in *The Metaphysics in Physics*, 2007. Oxford: Oxford University Press.

Maudlin, T. (2007a). *The Metaphysics in Physics*. Oxford: Oxford University Press.

Maudlin, T. (2007b). "What Could Be Objective About Probabilities?," *Studies in History and Philosophy of Modern Physics* **38**, 275–291.

Maudlin, T. (2011). "Three Roads to Objective Probability," in C. Beisbart and S. Hartmann (eds.), *Probabilities in Physics*. Oxford: Oxford University Press.

McCloskey, M. (1983). "Intuitive Physics" *Scientific American* **248** (4), 122–130.

Meheus, J. (2002). *Inconsistency in Science*. Dordrecht, Netherlands: Kluwer.

McDermott, M. (1995). "Redundant Causation," *The British Journal for the Philosophy of Science* **40**, 523–544.

Mellor, D. H. (1995). *The Facts of Causation*, New York: Routledge.

Menzies, P. (1989). "Probabilistic Causation and Causal Processes: A Critique of Lewis," *Philosophy of Science* **56**, 642–663.

Menzies, P. (1996). "Probabilistic Causality and the Pre-exemption Problem," *Mind* 105, 85–117.

Menzies, P. and Price, H. (1993). "Causation as a Secondary Quality," *The British Journal for the Philosophy of Science* 44, 187–203.

Mill, J. S. (1858). *A System of Logic: Ratiocinative and Inductive*, London: Longmans, Green and Co., 1930.

Ney, A. (2009). "Physical Causation and Difference-Making," *The British Journal for the Philosophy of Science* 60, 737–764.

Northcott, R. (2008). "Causation and Contrast Classes," *Philosophical Studies* 139 (1), 111–123.

Northcott, R. (2010). "Natural Born Determinists: a New Defense of Causation as Probability-raising," *Philosophical Studies* 150, 1–20.

Norton, J. (1987). "The Logical Inconsistency of the Old Quantum Theory of Black Body Radiation," *Philosophy of Science* 54, 327–350.

Norton, J. (2003). "Causation as Folk Science," Philosophers' Imprint 3 (4), http://www.philosophersimprint.org/003004/. Reprinted in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*, 2007. Oxford: Oxford University Press.

Norton, J. (2007). "Do the Causal Principles of Modern Physics Contradict Causal Anti-Fundamentalism?" in P. Machamer and G. Wolters (eds.), *Thinking about Causes: From Greek Philosophy to Modern Physics*. Pittsburgh: University of Pittsburgh Press.

Norton, J. (2008). "The Dome: An Unexpectedly Simple Failure of Determinism," *Philosophy of Science* 75, 786–798.

Nozick, R. (1969). "Newcomb's Problem and Two Principles of Choice," in N. Rescher (ed.), *Essays in Honor of Carl G. Hempel*, 114–146. Dordrecht: Reidel.

Paul, L. A. (2000). "Aspect Causation," *Journal of Philosophy* 97, 235–256, reprinted in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Paul, L. A. (2009) "Counterfactual Theories," in H. Beebee, C. Hitchcock, and P. Menzies (eds.), *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Price, H. (1991). "Agency and Probabilistic Causality," *The British Journal for the Philosophy of Science* 42, 157–176.

Price, H. (1996). *Time's Arrow and Archimedes' Point*. Oxford: Oxford University Press.

Price, H. and Corry, R., eds. (2007). *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press.

Price, H. and Weslake, B. (2009). "The Time-Asymmetry of Causation," in H. Beebee, C. Hitchcock, and P. Menzies (eds.), *The Oxford Handbook of Causation*. Oxford: Oxford University Press.

Putnam, H. (1975). "The Meaning of 'Meaning,'" in K. Gunderson (ed.), *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, VII. Minneapolis: University of Minnesota Press.

Ramachandran, M. (2004). "Indeterministic Causation and Varieties of Chance-raising," in P. Dowe and P. Noordhof (eds.), *Cause and Chance: Causation in an Indeterministic World*. London: Routledge.

Ramsey, F. (1928). "Universals of Law and of Fact," in *F. P. Ramsey: Philosophical Papers*, D. H. Mellor (ed.), Cambridge: Cambridge University Press, 1990, 140–144.

Reichenbach, H. (1956). *The Direction of Time*, Berkeley: University of California Press.

Roberts, J. (2009). *The Law Governed Universe*. Oxford: Oxford University Press.

Russell, B. (1913). "On The Notion of Cause," *Proceedings of the Aristotelian Society* 13, 1–26.

Russell, B. (1948). *Human Knowledge*. New York: Simon and Schuster.

Salmon, W. (1977). "An 'At-At' Theory of Causal Influence," *Philosophy of Science* 44, 215–224.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Salmon, W. (1993). "Causality: Production and Propagation," in E. Sosa and M. Tooley (eds.), *Causation*. Oxford: Oxford University Press.

Salmon, W. (1997). "Causality and Explanation: A Reply to Two Critiques," *Philosophy of Science* 64, 461–477.

Schaffer, J. (2000a). "Overlappings: Probability-Raising without Causation," *Australasian Journal of Philosophy* 78, 40–46.

Schaffer, J. (2000b). "Causation by Disconnection," *Philosophy of Science* 67, 285–300.

Schaffer, J. (2001). "Causes as Probability Raisers of Processes," *Journal of Philosophy* 98, 75–92.

Schaffer, J. (2003). "Is There a Fundamental Level?" *Noûs* 37, 498–517.

Schaffer, J. (2005). "Contrastive Causation," *The Philosophical Review* 114 (3), 327–358.

Schaffer, J. "The Metaphysics of Causation," *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition)*, Edward N. Zalta (ed.), URL = ⟨ http://plato.stanford.edu/archives/fall2008/entries/causation-metaphysics/ ⟩.

Sellars, W. (1962). "Philosophy and the Scientific Image of Man," in R. Colodny (ed.), *Frontiers of Science and Philosophy*. Pittsburgh: University of Pittsburgh Press. Reprinted in *Science, Perception and Reality*, 1963.

Skyrms, B. (1981). "The Prior Propensity Account of Subjunctive Conditionals," in W. L. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs*. Dordrecht: D. Reidel, 259–265.

Sklar, L. (1977). *Space, Time, and Spacetime*. Berkeley, CA: University of California Press.

Sober, E. (1985). "Two Concepts of Cause," in P. Asquith and P. Kitcher (eds.), *PSA 1984, vol. 2*, 405–424. East Lansing: Philosophy of Science Association.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge: MIT Press.

Stalnaker, R. (1968) "A Theory of Conditionals," in N. Rescher (ed), *Studies in Logical Theory, American Philosophical Quarterly Monograph Series, No. 2*, Oxford Basil Blackwell, 98–112, reprinted in E. Sosa (ed.), *Causation and Conditionals*, Oxford: Oxford University Press, 165–179, and in W. L. Harper, R. Stalnaker, and G. Pearce (eds.), *Ifs*, Dordrecht: D. Reidel, 107–128.

Steglich-Petersen, A. (2012). "Against the Contrastive Account of Singular Causation," *The British Journal for the Philosophy of Science* 63, 115–143.

Strevens, M. (1998). "Inferring Probabilities from Symmetries," *Noûs* 32, 231–246.

Strevens, M. (2003). *Bigger than Chaos: Understanding Chaos through Probability*. Cambridge, MA: Harvard University Press.

Strevens, M. (2011). "Probability Out Of Determinism," in C. Beisbart and S. Hartmann (eds.), *Probabilities in Physics*. Oxford: Oxford University Press.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Swoyer, C. (1982). "The Nature of Natural Laws," *Australasian Journal of Philosophy* 60 (3), 203–223.

Tahko, T. (2009). "The Law of Non-Contradiction as a Metaphysical Principle," *Australian Journal of Logic* 7, 32–47.

Talmy, L. (1988). "Force Dynamics in Language and Cognition," *Cognitive Science* **12**, 49–100.

Tooley, M. (1977). "The Nature of Laws," *Canadian Journal of Philosophy* 7 (4), 667–698.

Tumulka, R. (2006). "A Relativistic Version of the Ghirardi-Rimini-Weber Model," *Journal of Statistical Physics* **125** (4), 821–840.

Volchan, S. (2007). "Probability as Typicality," *Studies In History and Philosophy of Modern Physics* **38** (4), 801–814.

Wald, R. (1984). *General Relativity*. Chicago: The University of Chicago Press.

Ward, B. (2001). "Humeanism without Humean Supervenience: A Projectivist Account of Laws and Possibilities," *Philosophical Studies* **107** (3), 191–218.

Weslake, B. (2006). "Common Causes and the Direction of Causation," *Minds and Machines* **16** (3), 239–257.

Wolff, P. and Zettergren, M. (2002). "A Vector Model of Causal Meaning," *Proceedings of the twenty-fifth annual conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Wolff, P. (2007). "Representing Causation," *Journal of Experimental Psychology: General* **136** (1), 82–111.

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

Woodward, J. (2007). "Causation with a Human Face," in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press.

Woodward, J. (2012). "Causation and Manipulability", *The Stanford Encyclopedia of Philosophy (Winter 2012 Edition)*, Edward N. Zalta (ed.), URL = ⟨http://plato.stanford.edu/archives/win2012/entries/causation-mani/⟩.

von Wright, G. (1971). *Explanation and Understanding*. Ithaca, New York: Cornell University Press.

von Wright, G. (1974). *Causality and Determinism*. New York: Columbia University Press.

Wüthrich, C. (2011). "Can the World be Shown to be Indeterministic After All?" in C. Beisbart and S. Hartmann (eds.), *Probabilities in Physics*. Oxford: Oxford University Press.

Yablo, S. (1992). "Mental Causation," *The Philosophical Review* **101**, 245–280.

Yablo, S. (1997). "Wide Causation," *Philosophical Perspectives: Mind, Causation, and World* **11**, 251–281.

Yablo, S. (2002). "De Facto Dependence," *The Journal of Philosophy* **99** (3), 130–148.

{ INDEX }

abstreduction, 34–36, 123 of general causation, 36–37, 138–140, 186–187
acceleration, 88
actual causation, 43, 266
actual coarse-grained event, 62
advancement asymmetry, 16, 17, 226
agency, 143, 171, 198, 229, 253, 316
Albert, D., 250
amelioration of conflicts, 39–42
antecedent, 124
apparent conflict, 39–42
arbitrarily fast contribution, 88, 212
arena, 57 quantum-mechanical, 112–113 space-time, 79–82
aspect promotion, 154
asymmetry of advancement, 16, 17, 226 of bizarre coincidences, 184–186, 192, 239, 247–248 of causation, 105, 122, 230–239, 244–246, 249–262, 289–290 of entropy, 179, 244–247
asymmetry experiment, 226
attribute, 24

background of a contrastive event, 137
backtracking causation, 214 fixing, 203 influence, 214 nomic connection, 16, 203 prob-influence, 204, 207–210, 212, 214, 234–236 routes of influence, 214, 216, 240
backtracking experiment, 216–218
backtracking experiments B1, B2, B3, 240–242
Bell, J., 237
Bernoulli distribution, 181
bizarre coincidences, 184–186, 192, 233, 247–248 development, 129, 172 evolution, 129, 180, 181, 233 world, 129, 132, 192
Bohmian mechanics, 113–114, 116, 260
bold prediction, 228
Boltzmannian universe, 251–252
bottom conceptual layer of causation, 20, 22, 36, 37, 48–49, 53, 266, 278
brain correlator, 219–220, 240–242
branch systems, 244–246
broad prob-influence, 189–192, 237

c-connected, 96
c-path, 96
Cartesian spirit, 109, 219
Cartwright, N., 14, 122, 223, 274, 307
causal asymmetry, 105, 122, 230–239, 244–246, 249–262, 289–290 backtracking, 214 chain, 287, 300 contributor, 75, *see also* contribution culpability, 22, *see also* culpability eliminativism, 53–54 explanation, 18–19, 303 irreflexivity, 289 law, 53 line, 54 modeling, 19, 239, 314–316 pluralism, 197 process, 199, 297 relata, 21 variable, 316
causal contribution, *see* contribution
causal directness, 16, 17, 32, 49, 203–204, 234, 238, 239, 241 proof of, 207–211
causal process accounts, 3, 200
causation actual, 43, 266 by omission, 155 culpable, *see* culpability direction of, 105, 122, 230–239, 244–246, 249–262, 289–290 egalitarian, 46, 266 general, 22, 121, 140, 144, 284 metaphysics of, 13, 17, 19, 21, 44, 48, 152, 171, 194, 215, 225, 265, 277, 317 non-metaphysical aspects, 13, 17–20, 37, 44, 48, 267, 268, 306 singular, 22, 121, 265–266, 284
causation-like relation, 36
chain, causal, 287, 300
chaotic influence, 141
charge, 92, 93
classical gravitation, 56, 82–88 sparse interpretation, 85 standard interpretation, 82
classical mechanics, simple theory of, 25–26
classical unified field theory, 92–94
the clever, 235
closed time-like curves (CTCs), 111, 211, 229
coarse-grained event, 60, 62, 110 trivial, 61
coarse-graining, 201–202
Collins, J., 5
common-cause pattern, 207, 214–216, 223, 239, 242, 243
conceptual analysis, 4 empirical, xi, 1–20, 317 orthodox, 9–10
conceptual engineering, 3, 5, 8
conceptual insulation, 5
conditional, 123 contrary-to-fact, 123–124 corner, 125, 127 counterfactual, 123–125 nomic, 128–135, 138, 176, 251 ordinary language, 134–135
configuration space, 112
conflict, 39 amelioration, 39–42 apparent, 39–42, 302 genuine, 39–42
conflict amelioration, delegation of, 49, 267, 283, 302
conjunctive fork, 231, 254

consequent, 124
conserved quantity (CQ) accounts, 3, 6, 7, 200
conspiratorial coincidences, 87
content completeness, 93–94
content independence, 85, 94–98, 110
contextualization, 69 trivial, 69
contextualized event, 69
continuity of causal culpability, 303 of fixing, 161 of prob-influence, 161 of probability-fixing, 160 of terminance, 98–100, 211–212
contrary-to-fact conditional, 123–124
contrast, 136
contrastive causation, 155–157, 286
contrastive event, 136–138, 157 maximal, 137
contrastivity, 155–157
contrastivization irregular, 165 regular, 162
contribution, 75–76, 109, 121, 265–266 arbitrarily fast, 88, 212 instantaneous, 88 pure, 76, 77, 121 space-like pure, 88
contributor, 75, 109, *see also* contribution
control, 188, 250, 252
corner conditional, 125, 127
corpuscle, 25, 80
cotenability, 125
counterfactual, 123–124 mundane, 128, 130 vs contrary-to-fact conditional, 124
counterfactual dependence, 37, 123, 125–126, 135, 138
culpability, 22, 46, 266, 270, 272, 284, 285 as an explanatory device, 277 as proxy for promotion, 282, 284 for learning about promotion, 275 tension in concept of, 284 transitivity of, 160, 303
culpability$_1$, 286, 287, 291–292
culpability$_2$, 287, 292–299
culpability$_3$, 287, 299–302, 305
culpability$_4$, 287, 300–305
culpable causation, *see* culpability

Dainton, B., 245, 246
dappled worlds, 150
decision theory, 223, 243
dependence counterfactual, 37, 123, 125–126, 135, 138
derivative event, 60 future, 205 law, 66 past, 205 reality, 24–25
derivative quantity, 27
determinant, 67
determination, 23, 36, 67–68, 84, 110 by constitution, 77–78 trivial, 76 ubiquitous, 108, 176, 191
determinism, 84, 101–104, 108, 111 local, 101 possible-worlds, 102 unique-propagation, 102–103
deterministic law, 84, 102, 129, 173, 191
development, 129
difference-making, 37, 121, 135, 138, 142, 164
dimension, 80
direct cause, 314, 316
direction of causation, 105, 122, 230–239, 244–246, 249–262, 289–290
direction of influence, 142, 205–207, 257–262
team DO, 227
domain of contribution, 95
domain of dependence, 97 maximal, 103, 108
domain of influence, 95, 141
domain of terminance, 95, 97
team DONT, 227
Dowe, P., 2–4, 6, 200, 316, 317
Dummett, M., 230
duration, 80
dynamical law, 38

e-wards domain of influence, 98
Earman, J., 97, 102
effective strategies, 14–17, 170, 171, 197–199, 222–224, 226
egalitarian cause, 46, 266
electromagnetic charge, 93
electromagnetic field, 64, 89
electromagnetism, relativistic, 56, 89–92
eliminativism, causal, 53–54
emergence, 55
empirical analysis, xi, 1–20, 317 of the metaphysics of causation, 13–17, 44, 48, 152, 171, 194, 215, 225, 277, 317 of the non-metaphysical aspects of causation, 13, 17–20, 37, 44, 48, 267, 268, 306 two forms of, 12–13, 19–20, 48
Empirical Fundamentalism, xi, 31
enabling cause, 304
energy in ecology, 45 kinetic, 27, 30 mechanical, 28–30 thermal, 28–30
enion probability, 173, 174
entropy, 179, 184 asymmetry of, 179, 244–247
event, 58–64 actual, 59 coarse-grained, 60, 62, 110 contextualized, 69 contrastive, 136–138, 157 derivative, 60 fine-grained, 59 full, 63 fundamental, 59 impossible, 164 intermediate, 98 maximal, 70 maximal contrastive, 137 mundane, 46 possible, 59 temporally extended, 150–151 turnaround, 203 world, 64
Everettian quantum mechanics, 116
evolution, 129
example acid rain, 151 airbags, 153 anti-thermodynamic ocean, 245, 246 beach, 145 bear box, 159 being born, 228–229 bird perches, 182, 183, 237 blue roses, 189 bridge collapse, 288 cannonball in pond, 169, 170 college admissions, 194–196 cranberries, 154 dartboard, 172–173 duck reproduction, 304, 305 electrolysis of water, 202 falling rock, 172 fire-starter, 148 fizzled fuse, 277, 298 flag-raising, 220–222, 240–242 food/nutrient, 3–4 gas in box, 177 gas in box in tank, 177, 184 hungry dog, 155 intelligent plants, 230 itch and fever, 207, 239 light switch, 144 lightning, 146–147 lit panel button, 289 match striking, 287 melting

marshmallow, 158 moon, 141, 206 mouth, 188 nuclear-spin echo, 190–191 P-T extinction, 279–281 paint balls, 291–299 pizza, 228 power plants, 151 purple fire, 158 quark mass, 188–190 rock thrown in lake, 169 rock-throwers, 275 rust, 144 salty dessert, 155 salty dish, 156, 157 silver coin, 189 smoking gene, 242–243 spicy dish, 156 super-light cone, 212 thermostat, 229 thrombosis, 305 thunder, 147, 148 thwarting ubiquitous determination, 191–192 track switching, 300–304 tree falling on track, 301 volcanos and meteors, 279–281 voluntary stay in locked room, 261 wind farms, 151 windmill, 315
excision, 75
existent, 26–27
experiments B1, B2, B3 (backtracking), 240–242
experiments in empirical analysis, 11–12
explanation complete story, 17, 18, 271, 317 of effective strategies, 17
explicitly true, 7–8, 10, 48, 309

fibrillation, 178, 184
field electromagnetic, 89 gravitational, 86, 87, 91 physical, 80 strong, 92–93 weak, 92–93
Field, H., 307
Fisher, R., 242
fixed intermediate, 98, 160
fixing, 69–72 backtracking, 203 continuity of, 161 probability, 70–71 shielding of, 161 strong transitivity of, 158 trivial, 137 unidirectional transitivity of, 158–159 weak transitivity of, 157, 211
fizzle, 297–298 saved, 298, 318
flashy quantum mechanics, 115
flow of time, 70, 205
focal set, 287
foliation, 82, 91
food/nutrient example, 3–4
force dynamic model of causation, 285
foreground of a contrastive event, 137
fork asymmetry, 231, 255–257 conjunctive, 254 open, 254
forking model of counterfactuals, 258
fragment of history, 271
free will, 143
full cause, 67, 75
full event, 63
functional relations, 68
fundamental constants, 198 direction of influence, 142 event, 59 existence, 21 future, 205, 225, 257, 260 law, 66 level, 34 past, 205 reality, 24–25 volition, 109, 219
fundamental physics paradigm theories of, 56
fundamentally arbitrary parameter, 27, 33–35
future, 205
future-bizarre, 182–183
future-typicality, 182, 184, 233, 238, 256
fuzzing, 49, 59

Galilean space-time, 79, 82–83, 91, 112, 259
general causation, 22, 121, 140, 144, 284
general relativity, 56, 109–111
generic state, 84
genuine conflict, 39–42
global state, 81 of suitable shape, 101, 103
Godfrey-Smith, P., 11
Gold universe, 248–249
Goldstein, S., 181
Goodman, N., 124–132
graceful degradation, 5–6, 15, 187, 229, 251
gravitation classical theory of, 82–88
universality of, 94, 212
gravitational field, 86, 87, 91
gremlin, 108
growing block, 205
GRW interpretation of QM, 114

Hall, N., 5, 266, 300
Hitchcock, C., 10, 268
hole argument, 103
Horwich, P., 255
Humeanism, 65, 231

imposed contrastive effect, 296, 299
impossible event, 164
indeterminant, 72
indetermination, 72–73, 106
indeterminism Newtonian, 107–108
non-stochastic, 107–109 quantum, 114
stochastic, 104–106, 114
inducement, 285, 286, 294 successful, 285, 286
inextendible c-path, 97 space-like surface, 81
inferring prob-influence, 193, 196
influence, 122, 140–141, 167 backtracking, 214
chaotic, 141 direction of, 142, 205–207, 257–262 partial, 162–166 past-directed, 206
influence-based set of platitudes, 230
initial value problem, 109
insensitivity considerations, 172–174, 198
instantaneous causation, 87–88
instantaneous contribution, 88
interaction linkage, 93–94, 212
intermediate event, 98
intermediate region, 98
intervention, 239, 314–316
inus account of causation, 36–37, 76
inus condition, 76
irregular contrastivization, 164, 165, 222

kinetic energy, 27, 30
Kistler, M., 200

Lange, M., 31
lattice stochastic, 104–105
law, 64–67 derivative, 66 dynamical, 38
fundamental, 66 special science, 66–67
law facts, 64–66
law of non-contradiction, 38
law-like vs. accidental regularities, 16
level of reality, 34
Lewis, D., 65, 156

light cone, 90 of an event, 91
light-like, 90
link fuzzy, 187 precise, 187
local determinism, 101
locality, 166
locality, relativistic, 91
Locke, J., 261
Loewer, B., 250–252
Lorentz force law, 38
Lorentzian space-time, 109

Mackie, J. L., 36, 76, 311
macroperiodicity, 174
macroscopic condition, 179
macrostate, 177, 179, 251
managed inconsistency, 38, 45, 47, 283, 318
manifold, 79
manipulability accounts, 316
Markov process, 99
mass, 90
material contents, 57
material facts, 64–66
matter, 57
Maudlin, T., 181
maximal arena, 102 domain of dependence, 103, 108 event, 70 space-time, 101, 102
Maxwell's laws, 38, 65, 91
mechanical energy, 28–30
mechanism, causal, 199–201
metaphysics, 43, 267 of causation, 13, 17, 19, 21, 44, 48, 152, 171, 194, 215, 225, 265, 277, 317
metric Minkowski, 89 spatial, 83 spatio-temporal, 89 temporal, 80, 83
microcanonical distribution, 177, 181
microconstancy, 174
middle conceptual layer of causation, 20, 22, 36, 37, 49, 186, 266, 278
Mill, J. S., 36, 118, 124
minimal terminant, 84
Minkowski metric, 89 space-time, 79, 89–91
modal interpretations of QM, 116
morality and causation, 288
mundane counterfactual, 128, 130
mundane event, 46

naked singularity, 110–111
narrow prob-influence, 189–191, 237, 249
neuron diagram, 53
Newcomb's problem, 219
Newton's second law of motion, 84
Newtonian indeterminism, 107–108
Newtonian space-time, 259
Ney, A., 164
nomic conditional, 128–135, 138, 176, 251
nomic connection backtracking, 16
nomic connection, backtracking, 203
non-locality, 165
non-maximal space-time, 101
non-metaphysical aspects of causation, 13, 17–21, 37, 44, 48, 267, 268, 306
non-spatiality, 81, 212–213

Norton's dome, 108
Norton, J., 108
Nozick, R., 219
nuclear-spin echo, 190–191
null cone, 90
nutrient/food example, 3–4

$O = P$ prediction, 169–170, 190
objectivity, 23, 49, 54, 134, 268, 308 of causal direction, 249, 253–254 of probability distributions, 149, 174, 180, 181
omission, promotion by, 155
ontological vagueness, 150
open fork, 254
orthodox (conceptual) analysis, 9–10
overdetermination, 85–87

paradigm theories of fundamental physics, 56
parterminance, 72, 106, 115
partial cause, 75
partial influence, 162–166 space-like, 166
particle decay, toy theory of, 105–106
past, 205
past hypothesis, 185
past-bizarre, 184
past-directed influence, 206
pattern-based set of platitudes, 230
Paul, L. A., 5
perception, 281
phase space, 175
physicalism, xi, 55
pixelation, 179, 184
platitudes, 9, 10
point-like, 83, 89
possibility, 60
preëmption, 1, 150, 310–311, 318 early cutting, 298 late cutting, 298
Price, H., 233, 252–254
protrast, 136
prob-dependence, 122, 135–138, 140, 142
prob-influence, 140, 142–144 backtracking, 204, 207–210, 212, 214, 234–236 broad and narrow, 189–191, 237 continuity of, 161 pseudo-backtracking, 204, 222–223, 239–244 shielding of, 162
probability, 74, 274 enion, 173, 174
probability distribution, objective, 149, 174, 180, 181
probability-fixing, 23, 36, 70–71 continuity of, 160 shielding of, 161
probability-raising, 24, 149
problem of relevant conditions, 127–128
production, 37, 199
prominent background, 234
prominent foreground, 234, 236
promotion, 144–146, 152–153 aspect, 154 by omission, 155 contrastive, 155–157 trivial, 145
promotion experiment, 169–171
protrast, 160

pseudo-backtracking prob-influence, 204, 222–223, 239–244
psychology of causation, 17–18, 282–284
pure contribution, 76, 77, 121

quantum field theory, 56
quantum mechanics, 56, 112–116 Bohmian, 113–114, 116, 260 Everettian, 116 flashy, 115 GRW, 114 modal interpretations, 116 spontaneous collapse interpretations, 114–116, 166
quantum non-locality, 166
quantum state, 112
quinergy, 35–36

real basis, 25, 32, 33
real cause, 36, 67
realism, 308
realistic circumstance, 39
reality, 24
reduction, 34 abstreduction, 34–37, 123, 138–140, 186–187
reductive explanation, 34
region, 58
regular contrastivization, 162
regularities, law-like vs. accidental, 16
Reichenbach, H., 231, 254–255, 257
relationism, 58
relativistic electromagnetism, 56, 89–92
relativistic locality, 91, 166
relativity, 56
RELAXED standards of theoretical adequacy, 40–44, 302
routes of influence, 214 backtracking, 216, 240
Russell, B., 53–54, 68, 307

salient promoter, 287
Salmon, W., 200, 317
saved fizzle, 298, 318
Sellars, W., 8
settledness of the past, 262
shielding of fixing, 161 of prob-influence, 162 of probability-fixing, 161 of terminance, 99–100, 211–212
significant promotion, 290
simple theory of classical mechanics, 25–26
simplon, 77
Simpson's paradox, 152, 194–196
Simpson's reversal, 197
singular causation, 22, 121, 265–266, 284
singularity, 110–111
Sklar, L., 88
SM-conditional, 250–251
space invaders, 107
space-like, 83, 90 dimension, 80 partial influence, 166 pure contribution, 88 surface, 80 terminance, 81
space-time arena, 79–82, 213 connected, 79 Galilean, 79, 82–83, 91, 112, 259 homogeneous, 79 Lorentzian, 109 maximal, 101 Minkowski, 79, 89–91 Newtonian, 259
sparse interpretation of classical gravitation, 85
spatial metric, 83
spatio-temporal metric, 89
special science law, 66–67
spin echo experiment, 190–191
spontaneous localization, 114
standard interpretation of classical gravitation, 82, 83 of electromagnetism, 89
START, 226
state, 79, 81 generic, 84 global, 81
statistical mechanics, 175–180
stochastic lattice, 104–105
stochasticity, 69, 73, 74, 104–106
Strevens, M., 173–174
STRICT standards of theoretical adequacy, 40–44 for metaphysics, 43, 47, 48, 267
strong charge, 92
strong field, 92–93
strong transitivity of fixing, 158
stuff, 57
subevent, 63
subjective chance, 149
subjectivity, 308 of causal direction, 253–254
superevent, 63
superluminal influence, 91

temporal asymmetry of disjunctive deliberation, 252–253
temporal metric, 80, 83
temporal passage, 70, 205
terminance, 72, 116–117 and time, 81 continuity of, 98–100 in classical gravitation, 83–85 in relativistic electromagnetism, 91–92 intermediate, 98 minimal, 84 shielding of, 99–100 space-like, 81 transitivity of, 100 trivial, 76–78
terminant, 72, *see also* terminance
thermal energy, 28–30
thermodynamics, 175
time, 79, 81
time slice, 80, 83
time-like, 83, 90 dimension, 80
top conceptual layer of causation, 20, 22, 37, 50, 265, 269, 282
transference accounts, 3, 200
transitivity of causal culpability, 160, 303 of fixing, 157–159, 211 of terminance, 100
trivial contextualization, 69 determination, 76 promotion, 145 terminance, 76–78
Tumulka, R., 166
turnaround event, 203
typicality, 181–183

ubiquitous determination, 108, 176, 191
unidirectional transitivity of fixing, 158–159
unified field theory, classical, 92–94

vagueness, ontological, 150
volition, fundamental, 109, 219

wave function, 112

weak charge, 92, 93
weak field, 92–93
weak transitivity of fixing, 157, 211
Weslake, B., 252–254, 256
Woodward, J., 200, 314–317
world event, 64
world line, 80
wormhole, 100

Yablo, S., 289

Causation and Its Basis in Fundamental Physics

Bonus Chapters

Douglas Kutach

June 2013 Version

{ CONTENTS }

11. The Nomic Conditional and Natural Language Counterfactuals 1

11.1 Applicability to Singular Instances 4

11.2 Modus Ponens 4

11.3 Universal Modality 5

11.4 Centering 7

11.5 Actuality-Focusing 10

11.6 Antecedents as Underspecified Propositions 12

12. Morgenbesser's Coin 24

13. Orthodox Conceptual Analysis 35

13.1 The Orthodox Metaphysics of Causation 39

13.1.1 Causation and Ordinary Language Counterfactuals 41

13.1.2 Culpable Causes 42

13.2 The Orthodox Metaphysics of Culpable Causation 44

13.2.1 Close Fit to Psychological Data 44

13.2.2 Principled Analysis of Culpable Causes 49

13.2.3 STRICT *Standards for Accounts of Culpable Causes 49*

13.3 The Orthodox Metaphysics of Causation is Unneeded 51

13.4 Criticism of the Orthodox Metaphysics of Causation 52

13.4.1 Unclear Motivation 52

13.4.2 Causation is a Scientific Concept 54

13.5 Changing the Topic 55

13.6 Summary 55

14. The Psychology of Culpable Causation 57

14.1 The Toy Theory of Culpable Causation 59

14.2 Culpability$_1$ 62

14.2.1 Salience 62

14.2.2 Irreflexivity 64

14.2.3 Asymmetry 64

14.2.4 Significant Promotion 65

14.2.5 Causal Grouping Principles 66

14.3 Shortcomings of Culpability$_1$ 69

14.3.1 Precise Character of the Effect 70

14.3.2 Overlapping Causation 70

14.3.3 Probability-Lowering Causes 70

14.3.4 Trumping 71

14.4 Culpability$_2$ 72
14.5 Shortcomings of Culpability$_2$ 76
14.5.1 Saved Fizzles 77
14.5.2 Early Cutting Preëmption 78
14.5.3 Late Cutting Preëmption 78
14.6 Culpability$_3$ 78
14.7 Culpability$_4$ 79
14.8 Uncaused Events 85
14.9 Prevention 86
14.10Double Prevention 87
14.11Culpable Causation by Omission 88
14.12Summary 91

References 92

Index 96

{ 11 }

The Nomic Conditional and Natural Language Counterfactuals

This chapter contrasts my nomic conditional with popular alternative models of counterfactuals. The discussion is intended to motivate the conjecture that my nomic conditional is better suited to a scientific understanding of effective strategies than counterfactual conditionals that attempt to accord with natural language semantics. Such arguments are important for an empirical analysis of causation because what makes an empirical analysis successful in general is for it to provide a system of concepts that are optimized for explaining the empirical phenomena that motivate our having the folk concept. So for my empirical analysis of causation to be satisfactory, any counterfactual conditionals it employs need to be honed scientifically to fit nicely within a broader explanation of the phenomena that motivate us to believe in causation.

There are several limitations of the arguments offered in this chapter. Because I cannot possibly compare my account of counterfactuals to every logically possible account, I will restrict attention in this chapter to the more limited task of demonstrating that natural language counterfactuals are suboptimal for understanding influence and causation. Of course, a thorough evaluation of the relative merits of the nomic conditional can only be made by comparing the resulting account of causation to accounts founded on more traditional counterfactual semantics. Because the full consequences of the nomic conditional for causation are not spelled out until later chapters, the discussion in this chapter is necessarily incomplete. Nonetheless, the observations made here should help to further the thesis that the conceptual structures I engineered in the previous chapter are superior to standard models of counterfactuals for the purpose of clarifying the nature of causation.

Also, it is impossible to argue conclusively that models of counterfactuals based on natural language are necessarily inferior to the nomic conditional because there is no limit to the variety of modifications that can be made to improve the ability of natural language counterfactuals to address issues of causation. Furthermore, evaluating whether the nomic conditional is superior to other models of counterfactuals is not the kind of decision that can be made on the basis of clear, rigorous, explicit criteria. The arguments I offer rely somewhat on the reader's own discriminatory taste to identify whether a given feature of a counterfactual would make it less handy for explaining why some events are effective for bringing about certain effects. I hope that we share enough judgments to make progress in assessing the relative viability of natural language as a guide to an optimal measure of counterfactual dependence.

For readers who antecedently find it extremely implausible that the logic of natural language counterfactuals provides a promising theoretical structure for modeling influence and causation, all I can say is that the philosophical literature over the past hundred years has predominantly taken this hypothesis seriously and continues to do so to this day. In fact, I think it is fair to say that it is the received view.

A suitably neutral starting point for reasoning about hypothetical situations is to employ the following simple evaluation scheme: We first choose whatever situation we are interested in entertaining hypothetically, and then we let nature alone dictate what follows from that choice. The nomic conditional and natural language counterfactuals differ in how they make this scheme more precise.

With my nomic conditional, the scheme can be conceived as a two-step process. One first settles on a contextualized event $\overline{C}$ to represent the full situation one chooses to consider, typically a full event that extends far out into space so that it can be informative about events that occur minutes or hours or days later. Any vagueness or imprecision in one's hypothetical scenario is incorporated into $\overline{C}$ through the range of elements if includes and its probability measure. In step two, we let the fundamental laws of nature entail whatever they entail about $\overline{C}$. The result is some (typically larger) contextualized event, $\overline{G}$, which is defined as the largest contextualized event $\overline{C}$ fixes. The resulting $\overline{G}$ is a probability distribution over a set of fine-grained events. For any coarse-grained event E, either $\overline{G}$ fixes a probability for E, in which case that probability is the semantic value for $\overline{C} \mathrel{\Box\!\!\Rightarrow} E$, also equal to $p_{\overline{C}}(E)$, or it does not fix a probability for E, in which case the semantic value for $\overline{C} \mathrel{\Box\!\!\Rightarrow} E$ is undefined.

Notice that once $\overline{C}$ has been selected, no further material facts ever narrow or widen the scope of possibilities relevant to the semantic value of the counterfactual. That is, none of the possibilities encoded as elements of $\overline{G}$ are ever discarded as irrelevant to the semantic value of the counterfactual and no new possibilities are introduced at any later stage. Furthermore, nothing about the material layout of the actual world ever comes into play in the second step.

Although it is over-simplistic to make a similarly broad characterization about how to assess the semantic values for the kind of counterfactual conditionals formalized by philosophers, the following three-step procedure is not a bad first-order approximation. Step one involves settling on a proposition C to represent the hypothetical situation being entertained. It is virtually always the case that we intuitively have in mind a more narrow space of possibilities than the set of all worlds where C obtains, but strictly speaking, the content of the antecedent is equivalent to the content of the proposition C. In step two, we somehow combine information about C with information about the layout of material facts in the actual world and the laws of nature and additional contextual factors in order to arrive at a more narrow set of possibilities, which we may call the 'relevant C-worlds'. In similarity-based models of counterfactuals, these are the C-worlds that are most similar to actuality.[1] In context-based models, e.g. (18), these are identified as the contextually relevant C-worlds. In step three, we check to see whether E holds in all the relevant C-worlds. If it does, the counterfactual "If C were to occur, E" is declared true. If not, the counterfactual is declared false.

There are three main differences between my nomic conditional and natural language counterfactuals. First, the nomic conditional maintains a clean division between the content that is arbitrarily supplied as part of our free choice of which hypothetical situation we want to consider and the content that is supplied by the objective structure of nature. With the natural language counterfactual, the aspects that are arbitrary and the aspects that are objective are mingled in a complicated way. The arbitrary part comes partially through our specification of the antecedent C and partially through context-dependent parameters, e.g. a choice about the relevant notion of similarity to employ in order to arrive at the relevant C-worlds. The objective part comes partially through the laws of nature but also from accidents concerning how the material facts in the actual world are laid out, including future chance outcomes. The complicated character of how the arbitrary and objective aspects are mixed together is evident in several principles obeyed by the semantics—centering, actuality-focusing, and antecedents as underspecified propositions. I will explain these principles later in this chapter.

Second, the nomic conditional is especially handy for science because its semantic value is unadulterated by contingencies from the actual world that have no bearing on the effectiveness of strategies. Specifically, it does not take into account the actual future outcomes of fundamentally chancy processes. Natural language counterfactuals, by contrast, possess several channels through which accidental facts about the layout of material facts, such as fundamentally chancy outcomes, bear on the semantic values of counterfactuals and thus make them scientifically suboptimal.

Third, the nomic conditional is optimized for representing general relations of influence, i.e. what *kinds* of events are connected to each other in terms of probability-fixing relations. Unlike natural language counterfactuals, the nomic conditional makes no claims about what particular events influenced which other events on some particular occasion. See my discussion in §2.1.1 for a reminder of how coarse-grained events, including contextualized events, play the role of event types.

[1] Strictly speaking, similarity-based approaches can provide truth values for counterfactuals without a non-empty set of closest possible C-worlds. See discussion of the Limit Assumption in Lewis (40), p. 19–21.

The rest of this chapter is dedicated to clarifying how the nomic conditional is different from formalized versions of natural language counterfactuals vis-à-vis influence and causation. Along the way, I will suggest respects in which the nomic conditional is superior. Readers who are only interested in my positive account of causation or who find it highly implausible that the semantics of natural language conditionals could serve as a helpful constraint on one's theoretical machinery for modeling influence may be able to skip ahead to the next chapter without too much loss.

11.1 Applicability to Singular Instances

In every model of natural language counterfactuals, the kinds of resources that bear on the semantic value of a counterfactual, $C \mathrel{\Box\!\!\rightarrow} E$, depend significantly on whether C is true, and more narrowly on whether the relevant C-worlds happen to include the actual world. The privileged status of the actual world in the evaluation of natural language counterfactuals makes some sense given that the actual world has a privileged ontological status and thus deserves a special place in our interpretation of reality. In my account, however, for any contrastive event $\tilde{C} \equiv (\overline{C_1}, \overline{C_2})$, it makes no difference which of $\overline{C_1}$ and $\overline{C_2}$ is actual, and indeed it makes no difference whether they are both actual or both non-actual. One should recall that because we coarse-grain to get $\overline{C_1}$ and $\overline{C_2}$, we are in effect modeling *types* of fine-grained events, not tokens. Thus, the counterfactuals are representing *general* claims of influence, not claims about what a token event influenced on some particular occasion. This marks an important difference between the kind of counterfactual dependence in my account versus standard counterfactual accounts of causation which are based on evaluating single case influence.

One of the benefits of this approach is that it allows us to draw conclusions about counterfactual dependence in purely hypothetical circumstances in the same way we do for realistic or actual circumstances. We can evaluate, "If a unicorn were to have a sore leg, it would limp," by stipulating a $\overline{C_1}$ that represents some unicorn with a sore leg and a $\overline{C_2}$ that represents a healthy unicorn in the same background conditions, and then consult the fundamental laws to draw the conclusion that it would likely limp. It simply does not matter that both antecedents are non-actual.

11.2 Modus Ponens

A standard feature of most models of natural language conditionals, including counterfactuals, is that they obey modus ponens. Although counterexamples to modus ponens exist for natural language conditionals, it is one of the least controversial inference rules and is routinely incorporated into formal logic as a valid inference. Because the nomic conditional is probabilistic in character, it does not obey the kind of modus ponens rule that holds for truth-based models of counterfactuals. In particular, it is possible for $\overline{C}$ to fix a probability for E of one and yet there be a situation where $\overline{C}$ occurs and E fails to occur.[2] In this sense, the nomic conditional does not obey modus ponens. Because obeying modus ponens is arguably a defining feature of what it means to be a conditional, one could worry that the name 'nomic conditional' is misleading, but I think it is reasonable to conceive of the nomic conditional as a counterfactual conditional in the sense of being a formal structure for representing claims about what follows from hypothetical (typically non-actual) circumstances.

Modus ponens may be a reasonable principle for natural language counterfactuals that are intended to apply to individual events, but if the counterfactual is construed as a generality, i.e. concerning the general tendency of $\overline{C}$-type events to be followed by E-type events, then it makes sense only to hold a weakened version of modus ponens where the occurrence of $\overline{C}$ and the high probability of $\overline{C} \mathrel{\Box\!\!\Rightarrow} E$ is a defeasible reason to infer that E occurs.

It may be possible to formulate an adequate probability-based semantics for counterfactuals along the lines of Ernest Adams' (1)(2)(3) probability logic that does obey modus ponens, e.g. Skyrms (61), or Edgington (15). In Adams' logic, the validity of inferences tracks the preservation of high subjective probability. Non-conditional propositions are evaluated in terms of credence, one's subjective probability of their truth. Indicative conditionals are evaluated in terms of conditional subjective probability—the credence one has of the consequent conditional on truth of the antecedent—must also approach one. One could perhaps construct a parallel logic for my counterfactual model, by replacing Adams' talk of subjective probability with an appropriate notion of objective probability. If such an account could be consistently worked out, there would be a sense in which my nomic conditional does obey modus ponens, though certainly not the truth-preserving kind and certainly not in a way that completely matches ordinary intuitions about counterfactual inference. Terminology aside, refusing to alter my nomic conditional in order to respect intuitions about modus ponens leaves it insensitive to accidental circumstances of chance outcomes in the actual world, which provides a purer measure of how E depends in general on various contextualized events.

11.3 Universal Modality

A well known feature of natural language counterfactuals is that they (at least sometimes) incorporate some structure akin to a universal quantifier. When Jill says, "If the glass had fallen, it would have broken," one way that Jane can disagree is by responding, "No. What you said is incorrect. It *might* have survived." Jill's

[2] Of course, if $\overline{C}$ determines E and $\overline{C}$ is instantiated, E must occur.

counterfactual seems to mean something like, "If the glass had fallen, it definitely would have broken." Incorporation of some structure that resembles a universal quantifier is common to most theories of counterfactuals. It is present in Goodman's (19) account, Lewis's (40) account, Gauker's (18) account, and others. All such theories hold that counterfactuals of the form, "If *C* were to occur, *E* would occur," are true, roughly speaking, when *all* the relevant possible worlds where *C* holds, *E* also holds. For Goodman, the relevant possible *C*-worlds are the nomologically possible worlds with the appropriate (cotenable with *C*) background conditions. For Lewis, the relevant *C*-worlds are the *C*-worlds that are most similar to the actual world. For Gauker, the quantification ranges over linguistic contexts.

The problem with incorporating a universal quantifier is that there are virtually always at least some bizarre worlds where material processes radically disobey even the most reliable laws of the special sciences. Unless carefully crafted, this universal modality will result in the falsity of virtually every counterfactual dealing with mundane causal affairs. Although all theories of counterfactuals have some method to limit the scope of the quantifier, such restrictions are probably not sufficient to rule out enough bizarre worlds. Consider a person who held a hammer near a fragile vase at time *t* but did not strike it. Assume for the sake of example that the fundamental laws include ubiquitous stochasticity. Let *C* be 'the hammer struck the vase at *t*' and *E* be 'the vase broke', and consider the contrary-to-fact conditional, $C \mathrel{\Box\!\!\rightarrow} E$. In order for the relevant *C*-worlds to align sufficiently with the possibilities that we intuitively take to be relevant, they need to include all the worlds where the history matches actual history up until just before *t* and then evolves indeterministically and lawfully so that *C* occurs. Some theories place an additional constraint on these *C*-worlds in an effort to hold fixed some actual events of the future (see chapter 12), but these theories typically do not place restrictions on how fundamental chance outcomes turn out when they are causally immediately downstream from *C*. For example, if the vase had been struck, such theories might hold fixed the outcome of next week's lottery, but they will not hold fixed whether the vase breaks, whether the owner notices the shards, whether the owner cries, etc.[3] The problem is that—assuming all such worlds are among the relevant *C*-worlds—there will still be plenty of bizarre worlds where the hammer violently strikes the vase but where the molecules bounce around in some lucky way that keeps the vase intact. The problem in weeding out the bizarre worlds is that the only thing that makes them different from non-bizarre worlds is that their

[3] Lewis's (42) account advocates such a further restriction, so that some chance outcomes are counterfactually disallowed. It can turn out on his theory that in ordinary circumstances both of the following are true: "Had the vase been struck, it would not have broken," and "Had the vase been struck, it would have been overwhelmingly likely to break." Such results create a clear tension in the idea that counterfactuals incorporate an implicit 'definitely' in front of the consequent as noted above. More important, Lewis's theory allows biasing the outcomes of fundamentally chancy processes to match the gross character of the actual material layout, and his only reason for including the bias in his system is to accommodate some naive intuitions like those concerning Morgenbesser's coin, c.f., chapter 12.

future evolution is unusual. To rule them out just because of their unusual future behavior is hard to do without begging the question to some extent about what would have happened counterfactually.

One modification that could be attempted is to interpret consequents dealing with physical affairs as involving an implicit claim about chance. If the hammer had struck the vase, its chance of breaking would be very high. But this solution falls prey to the very same problem. Unless the fundamental laws are of a very unusual character, there will always be a few relevant worlds that assign a deviant chance. Because it appears to be very difficult to rule out bizarre worlds without creating worse problems, a better solution is to accept their existence and accommodate them. Intuitively, what one needs is some mechanism to smooth over a range of different C-worlds so that bizarre worlds are included but assigned fantastically small probability and so that if different C-worlds have different chances for E, there can still be a single effective chance for E. Contextualized events!

11.4 Centering

For conditionals evaluated according to a truth-based semantics, the validity of modus ponens together with the validity of the inference rule, “From $\alpha \& \beta$, infer $\alpha \mathrel{\Box\!\!\rightarrow} \beta$,” implies that when the antecedent is true, the truth value of the counterfactual is equal to the truth value of the consequent. Applied to counterfactuals expressing influence, this means that if one starts out intending to evaluate the nomic consequences of a $\overline{C}$ that is actually instantiated, then (in order to obey truth-based modus ponens), the value of “If $\overline{C}$ were to occur, then E,” must come out equivalent to the truth value of E. In Lewis's (40) account of counterfactual logic, the semantic principle underlying this reasoning is labeled ‘centering,’ and it can be thought of in terms of comparative similarity as follows. There is one world that is most similar to the actual world: the actual world itself. One can extend the basic idea behind centering to alternative counterfactual logics as follows: Centering holds if and only if whenever α holds, $(\alpha \mathrel{\Box\!\!\rightarrow} \beta) \equiv \beta$ holds. (‘$\equiv$’ here represents the truth-functional biconditional.)

Centering is a controversial principle insofar as it is intended as a general principle governing natural language counterfactuals.[4] Since our discussion concerns influence, however, we only need to consider a weaker, potentially less controversial version of centering. Counterfactuals relevant to relations of counterfactual dependence only have antecedents and consequents that proclaim the existence

[4] The most common objection is that the argument form, “From $\alpha \& \beta$, infer $\alpha \mathrel{\Box\!\!\rightarrow} \beta$,” allows one to take any two random truths and infer a counterfactual relation between them. Philosophers have a standard response, which is to argue that the inference is technically valid but that in natural language, such an inference carries an implicature that is not represented in the formal system. The seemingly invalid inferences turn out to be just those inferences that carry a misleading implicature. The inferences are always valid but can sometimes be misleading.

or non-existence of some event, so we only need to consider centering insofar as it concerns this more limited class of counterfactual conditionals.

Centering is assumed in David Lewis's (40) account of counterfactual dependence among events, which is based on a model of events where they are both coarse-grained and singular, i.e. where token events need not be (and typically are not) maximally fine-grained. Counterfactual dependence of some E on some C is thus interpreted as a dependence between those two particular events in a single instance. In particular, an actual event E counterfactually depends on an actual event C if and only if $(C \;\Box\!\!\rightarrow\; E)\&(\neg C \;\Box\!\!\rightarrow\; \neg E)$. In virtue of centering, the formula reduces: Counterfactual dependence exists if and only if $\neg C \;\Box\!\!\rightarrow\; \neg E$.

Within the context of an empirical analysis, the goal is to identify a notion of counterfactual dependence that applies to general relations of influence: whether events of type E counterfactually depend on events of type C. So, Lewis's truth conditions for counterfactual dependence are not directly relevant to the issue at hand. For example, if E is a fundamentally chancy event, say a particle decay that occurs after C, and the presence of C does not raise or lower the chance of E, then we ought to say that C does not influence E in the sense relevant for modeling effective strategies. Hence, if we want a tight connection between influence and counterfactual dependence, we should say that E does not counterfactually depend on C. But on a reasonable construal of Lewis's account, E does depend on C because if C had not occurred, the events of the future would have evolved according to laws that do not hold fixed E's future occurrence. Lewis (42) complicates this assessment by suggesting that some future events like E should be held counterfactually fixed when altering C, but no principle is ever provided to guide our assessment of which future events should be held fixed. This example is not a counterexample to Lewis's account because he is not addressing how events in general depend on one another; it only demonstrates that Lewis's construal of counterfactual dependence will not provide the sought-after relation. I will not digress any further here to investigate whether some alternative account of singular counterfactual dependence could, in a more roundabout way, provide a foundation for an account of general counterfactual dependence useful for understanding causation.

If one were to attempt to shoehorn my nomic conditional into obeying centering, $\overline{C} \;\Box\!\!\Rightarrow\; E$ would presumably need to be semantically disjunctive. Its value would be $p_{\overline{C}}(E)$ when $\overline{C}$ does not actually occur and either 1 (if E is true) or 0 (if E is untrue) when $\overline{C}$ does actually occur. The resulting measure of counterfactual dependence would reduce to $1 - p_{\overline{\neg C}}(E)$ when E is true, and $0 - p_{\overline{\neg C}}(E)$ when E is false. Let 'influence$_b$' be the label for the degree of prob-influence corresponding to these two formulas, and let us call it the 'bifurcated model' of counterfactual dependence. For brevity, we can just focus on the special case where the antecedent and consequent are actual events, an event c and a later event e (coarse-grained as E). In order evaluate influence, we need to consider c insofar as it is part of some state c' that is an extension of c that is big enough to termine e, and in order to focus on the influence of the very particular way this extension is instantiated, we should contextualize c as the trivial contextualization, $\overline{C}$, that contains only the single element, c'. For the special case where the antecedent is true, the degree to which c influences$_b$ E is equal to $1 - p_{\overline{\neg C}}(E)$.

When the laws are deterministic, influence$_b$ approximates an adequate representation of influence in many circumstances.[5] For example, when c does not contribute to E because, say, it is outside of E's past light cone, then c does not influence$_b$ E. To give another example, my waving a hand does not significantly influence$_b$ the position of the moon one second later because given any plausible alternative activity one could engaged in, the moon very probably would be located very near its actual position. Influence$_b$ also does a reasonably good job of measuring influence when E would have been made improbable by c's nonoccurrence. For example, when Guy wins a raffle, it is reasonable to say that the precise number of rotations of the ticket barrel influenced whether Guy won because had the barrel been rotated one more or one less time, Guy probably would not have won. In all these cases, influence$_b$ tracks our intuitive grasp of influence.

What makes influence$_b$ a poor guide to influence generally is its inability to properly handle stochastic dynamical laws. Suppose that c is the actual flapping of a certain butterfly's wings and e (coarse-grained as E) is an actual lightning strike at a fairly specific location in the sky at a fairly specific time in the distant future, say to within a few seconds. Furthermore, suppose that the fundamental dynamics is extremely chaotic. Owing to the chaos and the rarity of lightning strikes, $p_{\overline{\neg C}}(E)$ is very low; hence, the butterfly strongly influences$_b$ the lightning. This verdict accords with the intuition that had the butterfly done something else, there would likely not have been a lightning strike at that specific time and place. But its influence$_b$ does not reflect that the butterfly was an insignificant contributor to E. We know that no matter what the butterfly did, lightning probably would not have struck. Assuming E's great sensitivity to the many chance processes during the intervening years, replacing the butterfly's instantiation with any other remotely reasonable fine-grained event will lead to very nearly the same probability of E. The many chance outcomes magnified through numerous chaotic microscopic interactions drown out the butterfly's probabilistic contribution.

The lesson to be drawn for my account, I think, is that my measure for counterfactual dependence, $p_{\overline{C}}(E) - p_{\overline{\neg C}}(E)$ should not be refigured to accommodate centering by adopting the bifurcated model of counterfactual dependence.

In any case, by virtue of the controversial nature of centering, it proves useful to examine problems with models of natural language counterfactuals that arise when centering is abandoned and a weaker principle is maintained. The weaker principle that is of primary interest is actuality-focusing.

[5] Recall that I use the general term 'influence' to stand for the imprecise collection of all reasonable notions of influence, including our instinctive grasp of influence. 'Influence' is not a technical term in my system.

11.5 Actuality-Focusing

Suppose I have a barrel for collecting rainwater, and I think it was empty yesterday. When I say, "If the barrel had been full yesterday, I would have poured its contents on my garden," that seems to express a truth that is implied by my desire to water my garden, my awareness and ability, the lack of anything that would prevent my watering the garden with the rain barrel, etc. If anyone points out that my counterfactual is false because the barrel could have been full of mercury (which I would have noticed and not poured on my garden), I could rightly respond that that possibility is not within the range of contextually relevant possibilities I was discussing, i.e. not instantiated by any event in the antecedent, $\overline{C}$, that I was intending to communicate and so is irrelevant to the correctness of my assertion. But suppose, by some odd happenstance, that my rain barrel was actually full of mercury. By ordinary standards, my stated counterfactual was incorrect because, as a rule, if the actual state makes the antecedent true, then the space of possibilities I thought was relevant to the correctness of my counterfactual—those that instantiate a barrel full of water—are rendered obsolete; the only relevant possibilities are those that instantiate the actual state with a barrel full of mercury. The practice we have of evaluating counterfactuals by overriding the imagined space of relevant possible states with the actual state when the antecedent is true may be denoted ***actuality-focusing***.

For illustration, consider a counterfactual $C \mathrel{\Box\!\!\rightarrow} E$ where C expresses the occurrence of some event at time t and E expresses the occurrence of some later event. Actuality-focusing claims that if C is true, the space of relevant C-worlds only contains worlds that instantiate the actual state at the time of C's occurrence. Actuality-focusing is weaker than centering because it does not enforce a rule that what happens (counterfactually) after C must match what happens in the actual world. In the framework of my account, the nomic conditional is evaluated by starting with some sufficiently filled out $\overline{C}$ and letting the laws evolve that state forward in time. If actuality-focusing were imposed on my account, it would say that whenever the actual state $S_{@}$ instantiates $\overline{C}$, one should evolve $S_{@}$ forward in time instead of $\overline{C}$.

Actuality-focusing imposes two complications. The first is that it allows contingent circumstances in the material content of the universe to override the scenario that was intended by the antecedent. In the example above, the intended subject of discussion concerned what would have likely followed from having a barrel full of *water*. The surprising actual circumstances meant that the truth value of the spoken counterfactual did not have anything to do with the intended subject of discussion. If the counterfactual had explicitly mentioned the barrel being full of water in the antecedent, however, the truth value of the spoken counterfactual would have addressed the intended subject. I will postpone further discussion of this complication to §11.6 because it is a special case of the broader problem of mismatch between the intended content and the content rendered by the model of counterfactuals.

Second, actuality-focusing forces us to use only the fine-grained actual state for the antecedent state when the antecedent proposition is true instead of allowing coarse-grained events like the nomic conditional does. Insofar as we want to employ natural language counterfactuals for understanding effective strategies, the forced fine-graining creates several suboptimalities.

First, because of volitional limitations, we are unable to employ fine-grained events in practice. We cannot control our actions perfectly precisely; at best we can reliably create some coarse-grained event C while making some ways of instantiating C more likely than others.

Second, we only have epistemological access to a limited fraction of the coarse-grained states and virtually no access to fine-grained states. So if we are to learn about the causal regularities it will come by way of learning about influence insofar as it involves coarse-grained events. Furthermore, because strategies themselves are coarse-grained, (as discussed in §5.1), we need to be able to make assessments of counterfactuals involving coarse-grained events in order to assess the effectiveness of strategies.

Third, if the laws are deterministic, actuality-focusing reduces prob-dependence to the problematic bifurcated notion of counterfactual dependence discussed in §11.4. To see how the problem arises, recall the example where c is the flapping of a certain butterfly's wings and e is a future lightning strike in the distant future, coarse-grained as E. According to my notion of counterfactual dependence—prob-dependence—we are free to construe the flapping in a coarse-grained way as the contextualized event $\overline{C}$ and the contrast as a similarly contextualized non-flapping, $\overline{\neg C}$. The prob-dependence of E on the flapping, $\tilde{C}$, is the probability of E given the flapping (with its environment) $\overline{C}$ rather than non-flapping (with its environment) $\overline{\neg C}$. Let us call the prob-dependence of E on $\tilde{C}$ the 'coarse-grained influence' of the flapping on the lightning. Its value is $p_{\overline{C}}(E) - p_{\overline{\neg C}}(E)$. If we impose actuality-focusing, we must instead evaluate the first counterfactual as the probability of E fixed by some actual state $S_{@}$ at the time c occurs. Let us call the corresponding degree of influence the 'fine-grained influence' of the flapping on the lightning. Its value is $p_{S_{@}}(E) - p_{\overline{\neg C}}(E)$. If the fundamental laws incorporate enough stochasticity, the difference between fine-grained and coarse-grained influence will be insignificant. But when the fundamental laws are deterministic, $p_{S_{@}}(E) = 1$ whereas $p_{\overline{C}}(E)$ is nearly zero. So, under determinism, the flapping counts as strongly influencing the lightning in the fine-grained sense but insignificantly influencing the lighting in the coarse-grained sense. We can easily make sense of this discrepancy: The mere fact that the butterfly flapped some way or other did not alter the chance of lightning, but that it flapped in the very particular way it did, in the very particular environment it was in, did greatly influence the lightning. Had it not flapped that way, there almost certainly would not have been a bolt of lightning ten years later at the assigned location. Fine-grained influence is not an illegitimate notion of influence, nor is there anything wrong with having a discrepancy between the fine-grained and coarse-grained influence. However, if we are interested in influence insofar as it bears on effective strategies, we ought to resist having our measure of influence among actual events *forced* into being only the fine-grained kind. The coarse-grained measure of influence is useful for measuring what the flapping of butterflies affects in general. The fine-grained influence is only applicable to the vast precise microstate $S_{@}$ and because of its practical epistemic inaccessibility and our practical inability to reproduce $S_{@}$, it does not do us much good. So, by refusing to impose the actuality-focusing that natural language counterfactuals incorporate, we leave ourselves free to use the fine-grained influence if we wish or ignore it in favor of coarse-grained influence.

11.6 Antecedents as Underspecified Propositions

In the philosophical literature, natural language counterfactuals are typically modeled as a connective between two propositions. When the counterfactual is mundane—i.e. dealing with an antecedent and consequent that each can be interpreted as expressing the existence or non-existence of some mundane event—the underlying semantics of the counterfactual incorporates the antecedent event only insofar as it can be represented by a proposition. What's more, when we discuss a contrary-to-fact situation by specifying a proposition, we typically only communicate a limited number of salient facts about the intended situation, often by presuming that other facts are to be drawn from the layout of the actual world. If our only interest were to communicate information about the objective probability of event E given the non-actual contextualized event $\overline{C}$ and brevity were no consideration, one could just say that $\overline{C}$ makes E have probability p. Because it takes too long to communicate the content of $\overline{C}$ explicitly ($\overline{C}$ being very big and detailed) and because we are almost always interested in contrary-to-fact situations that closely resemble actual states and don't depend on the precise details of distant events, it is convenient for us just to mention the localized event C with the understanding that the full intended counterfactual situation can be reconstructed by taking the actual state and altering it appropriately to make C obtain.

Restricting our attention to mundane counterfactual statements from here on, let us say that a ***consequent event*** is just the event capturing the content of the consequent (which can be any size) and an ***antecedent event*** is an event capturing the content of the antecedent but big enough and filled out enough to fix what happens at the location of the consequent event. The antecedent event is virtually always spatially bigger than anything explicitly cited in a mundane counterfactual statement. Let us say that an ***underspecified antecedent*** is a propositional antecedent that does not specify the antecedent event richly enough for it to termine the consequent event. In practice, natural language counterfactuals virtually always employ underspecified antecedents. In principle, though, one could specify a proposition that is rich enough to express an antecedent event, in which case it would count as a ***sufficiently specified antecedent***.

The difference between an underspecified and a sufficiently specified antecedent consists in what resources need to be brought to bear in order to flesh out enough of the content of the explicitly stated antecedent for it to be useful for understanding influence. A sufficiently specified antecedent only needs one kind of resource: clarification of the intended meaning of the explicitly stated antecedent. The resources of this kind include (1) making the extension of the antecedent precise while remaining consistent with the speaker's intentions and the context, (2) identifying the referents of any pronouns or demonstratives or implicit indexicals, and (3) accommodating any semantically loaded terminology that would unduly bias assessments of influence. An underspecified antecedent, by contrast, needs more than just a clarification of the meaning of the explicitly stated antecedent to be brought to bear on issues of influence; it requires adding facts about the material layout of the actual world.

Consider the following example, which is representative of most counterfactuals involving influence. In the actual world, Jane and Jill are standing near a large bucket placed on a shelf high enough up so that no one can see inside. Neither has information about the bucket's contents beyond standard background knowledge. Jane says, "If I were to toss a dry ball into this bucket, it would become wet." Jill disagrees. Without ever throwing a ball to test the conditional, they look in the bucket and find it empty and dry. By all ordinary standards, that definitively settles the dispute in Jill's favor, assuming there are no other relevant causal mechanisms being left out of the description.

Notice the following three points. First, there is nothing in the meaning of the antecedent, no matter how it is unpacked and clarified, that has anything to do with settling whether the bucket was dry in the counterfactual situation being entertained. If Jane had said afterward, "More precisely, I meant 'If I were to toss a ball into this bucket under circumstances where the janitor had previously filled the bucket with water, the ball would become wet,' " that would count as changing the topic. By all accounts, the dispute concerned the character of the bucket as it was when the statement was made. Second, the fact that the bucket was actually dry and was recognizable as such played an ineliminable role in the counterfactual's being demonstrably false. Third, any account of counterfactuals where Jane's claim does not come out false (or very improbable), and more generally fails to accord with most simple cases of counterfactuals dealing with interactions among physical objects, would be suspect because it would be unclear how that account would be relevant to influence. Although it is incorrect to dismiss an account of counterfactuals for not matching naive intuitions, in the particular subset of cases where the facts about influence are straightforward, a counterfactual intended to optimize talk of influence ought to make clear sense of them even if it does not render them explicitly true.

Setting theory aside, the example of the bucket shows that our natural reading of the counterfactual relies on facts concerning the actual layout of history in order to fill out the antecedent event. Thus, natural language counterfactuals employ underspecified antecedents. By contrast, no such deference to the actual material layout is built into my nomic conditional. There are instead just different antecedents one can consider. One can let $\overline{C_w}$ be a contextualized event instantiating Jane tossing the ball into a water-filled bucket and let $\overline{\neg C_w}$ be the contextualized event just like $\overline{C_w}$ except that Jane does not toss the ball. The contrastive event $\tilde{C}_w$ is then just defined to be this ordered pair of contextualized events, whence the fundamental laws entail that the ball is much more likely to become wet. One is also free to consider the contrastive event $\tilde{C}_d$, which is stipulated to be the ordered pair consisting of the contextualized event $\overline{C_d}$ where the bucket is dry with Jane tossing the ball in and the contextualized event $\overline{\neg C_d}$ that is identical except that Jane is not tossing the ball. The fundamental laws imply that $\tilde{C}_d$ does not make the ball any more likely to become wet. One is also free to consider a contrastive event that represents a weighted mixture of the two possibilities, or a contrastive event like $(\overline{C_w}, \overline{\neg C_d})$. All these conditionals can be considered, with each representing a different relation of counterfactual dependence. None of them, though, are directed at representing the claim that Jane and Jill were disputing. In order to represent that claim, one would need to characterize the contrastive event not by specifying its condition directly in terms of various fine-grained events, but indirectly by specifying its elements as modifications to whatever state of the world exists at the time Jill and Jane are debating. Within the context of my account, one can make sense of their debate as follows. Jane is saying in effect that the actual state at t is such that if it were naturally coarse-grained to include her throwing the ball into the bucket, the ball would probably become wet. Looking in the bucket after t provides evidence that the actual state at t had a dry bucket, which would be contextualized into a dry-bucket-instantiating event like $\overline{C_d}$, which fixes a low probability of the ball becoming wet.

The difference between underspecified and sufficiently specified antecedents can be summarized, as in Fig. 11.1, in terms of how the evaluation of counterfactuals would proceed in each case. Their essential difference consists in how facts about the actual material layout are accommodated, e.g. the fact that the bucket was actually dry. In accounts that only use sufficiently specified antecedents, such as mine, facts about the actual material layout come into play only as factors one is free to consider when deciding on the antecedent event, i.e. which hypothetical situation one chooses to consider. The actual material layout enters at the same step as decisions about how to fix the extension of the antecedent event using the explicitly stated antecedent, context, speaker's intentions, etc. In accounts that use underspecified antecedents, such as any account based on world-similarity, the actual material layout also comes into play after one has completely settled on what hypothetical one chooses to consider. Even after using the explicitly stated antecedent, context, speaker's intentions, etc., to fix the extension of the (propo-

	Sufficiently Specified Antecedents	Underspecified Antecedents
Step One (Stipulative Part)	Fix an antecedent event $\overline{C}$ and the extension of E.	Fix the extension of C and E and any theoretical parameters (e.g. a choice of similarity relation).
Step Two (Substantive Part)	Consult the laws.	Consult the laws, the actual material layout, and the theory of counterfactuals.

TABLE 11.1 Two general frameworks for evaluating mundane counterfactuals.

sitional) antecedent, there are further facts about the actual material layout that come into play to determine the correctness of the counterfactual, including primarily any background conditions transferred from the actual world to the full counterfactual situation.[6]

The purpose of this section is to show that for the purposes of an empirical analysis of causation, models of counterfactuals that employ underspecified antecedents are inferior to models that employ sufficiently specified antecedents. Since models of natural language counterfactuals use underspecified antecedents, they will count as inferior (on this issue) to my nomic conditional, which uses sufficiently specified antecedents.

As the bucket example demonstrates, the ordinary way of evaluating counterfactual claims follows the 'underspecified antecedent' way of evaluating counterfactuals because in the first step we translate "I toss a ball into this bucket" into the proposition that Jane tosses a ball into the bucket that her intention selects. No part of the content of that antecedent bears on whether the bucket to be considered in the counterfactual reasoning is dry. At best, there exists an implicit indexical reference to actuality, as in "Jane tosses a ball into the bucket that her intention selects and everything else going on is like it is in Jane's actual environment." Whether the bucket in the counterfactual scenario is dry is only fixed afterward, when one's theory of counterfactuals is consulted to identify the semantic value of the counterfactual by factoring into account the actual conditions and the antecedent proposition.

Accounts that only use sufficiently specified antecedents can make some sense of the implicit indexical reference to actuality in natural language by conceiving of those counterfactuals as incorporating a map in the sense employed by computer programmers. The kind of map relevant to counterfactual evaluation is an index of all the possible states at time t, linked to a state at t that has been modified appropriately to instantiate the antecedent event. To evaluate 'Jane tosses a ball into the kind of bucket her intention selects while everything else going on is like it is in

[6] There is no requirement that a theory actually use all the resources listed in the table, only that they are in general free to use them.

the actual world at time *t*,' one would try to reckon, as best one can, the actual state of the world at *t*, use the map to figure out what the modified state is, and then let the laws operate on that modified state. The map structure in effect allows us to take the indexical reference to actuality out of the semantics of the counterfactual and instead impose it as an external parameter that fixes which antecedent event is relevant given the actual material layout. Call this the 'map approach.'

In order to be useful for understanding effective strategies, the semantic values of counterfactuals need to cohere with the common practice of using laws of nature to infer from causes to effects. I take as a starting point that whatever comes out of a theory of how to evaluate counterfactuals, if it employs underspecified antecedents, its pronouncements regarding mundane cases of influence ought to match something in the ballpark of the following ***minimal account of counterfactuals***:

1. One takes the actual state of the world at a time *t* pertaining to the antecedent and modifies it appropriately to instantiate the antecedent.
2. One lets the appropriate laws evolve that state into the future.
3. One looks to see whether the consequent is entailed by that future evolution (or at least made probable by the counterfactual state).
4. The semantic value of the counterfactual is just the semantic value of the consequent in the alternative evolution of history. This could be the truth value of the consequent or perhaps the probability of its truth.

I mean to include all of the following variants as "in the ballpark":

- Accounts that hold counterfactually fixed some actual facts after *t* or make the counterfactual evolution more likely to match actual facts than the laws indicate[7]
- Accounts that use special science laws, folk psychological principles, or other similar rules of thumb for generating the counterfactual historical development
- Accounts that also evolve the modified actual state into the past
- Accounts that permit counterfactual backtracking, e.g. counterfactual inferences are made toward the past and then toward the future.
- Forking accounts
- Extended forking accounts

A forking account is an account of counterfactuals where the relevant possible worlds all match the history of the actual world up until some time t' not too soon before *t*. At t', the indeterministic evolution of the counterfactual history departs (or forks) from actuality and leads lawfully and more or less naturally to

[7] Even though some actual post-*t* facts are carried over to the counterfactual history, the rest of the evolution should come from the laws.

the antecedent obtaining at *t*, and then later lawfully to the rest of the counterfactual history. An extended forking account generalizes the applicability of forking accounts to deterministic settings by permitting miracles to generate the fork.

An acceptable theory does not need to model counterfactuals in a way that literally follows the above steps, but it ought to have its end results match near enough the results one gets from the above procedure. For example, Goodman's hoped-for theory would have fit the bill. The justification for this condition of adequacy is merely that otherwise it is hard to make sense of our practices regarding verification and falsification of counterfactuals. There are only two kinds of tests that provide direct evidence about the correctness of some chosen mundane counterfactual. (Indirect evidence could come by way of being evidence for or against some alleged law of nature.) The first kind of test is to find out that the antecedent is true, in which case the counterfactual's semantic value tracks that of the consequent. The second kind of test is to conduct an experiment where one's initial conditions are just as in the actual world except modified to make the antecedent true. The results generated by these two kinds of tests are not generally sufficient to establish the semantic value of a counterfactual. The first kind of test does not apply to contrary-to-fact conditionals, and the second is only loosely applicable because what happens in an actual test of some other initial conditions somewhere else does not necessarily indicate what would have happened in the single instance pertaining to the counterfactual.

Ultimately, claims about what would have happened in this one region *R*, had things been otherwise, are epistemically inaccessible. Our only grip on them comes by way of the conceptual connections among counterfactual conditionals, truth, probability, laws, material facts, time, and so on. In general, one can identify many notions of counterfactual that do not match the results of the above steps. Some mismatches are a result of the counterfactuals being used to express semantic or logical relations. Others are idiomatic: "If I were you,…." Others express epistemic relations. Even within the context of counterfactuals that seemingly express influence, one could cook up ridiculous unfalsifiable rules, e.g. that any contrary-to-fact conditional whose antecedent refers to gold is true if its consequent refers to goblins. My declaration that an adequate theory of counterfactual evaluation must match the results of the above procedure is intended to rule out theories that are too remote from our implicit practices for checking counterfactual claims to count as 'optimizing a notion of counterfactual toward the explanation of effective strategies.'

The central problem to be solved by any account of counterfactuals that employs underspecified antecedents and purports to be optimized for understanding influence and causation is how to get from the underspecified antecedent to something approaching the results of the minimal account. There are two ways of approaching this problem. The first, called 'the informal approach,' involves just relying on our implicit ability to fill out the underspecified antecedent as needed in order to match what intuitively seems like the correct (fleshed out) antecedent event. It is not always transparently clear whose theories are intended to accord with which approach, but I think it is fair to associate the informal approach with a wide variety of theories, e.g. (5)(6)(16)(48).

The second way, called 'the principled approach', is to provide a principled theory that dictates how to narrow the space of possibilities permitted by the underspecified antecedent so as to arrive at a semantic value for the mundane counterfactual. The point of the principled theory is to provide a scientific replacement for our intuitive pragmatic grasp of the contextually relevant background conditions. Remember that for mundane counterfactuals, the mere truth of the antecedent virtually never suffices to fix whether the consequent obtains (or even a probability for the consequent). The principled theory tries to identify a much smaller correct set of relevant possibilities that does suffice to inform us of the consequent. Goodman's theory, I think, is aimed at this principled approach, but he never actually provided a theory of how to restrict cotenability, so he never offered a viable principled theory. One principled theory for handling underspecified antecedents is the forking model, where the relevant worlds are those that exactly match the material layout of the actual world previous to some designated forking time t_f and obey the actual laws thereafter, and have the antecedent obtain.[8] Another principled theory is Lewis's (42) theory of world-similarity.

In the rest of this section, I intend to demonstrate that both approaches are suboptimal for understanding causation. If correct, that will favor my account, which is an alternative to what these two models of counterfactuals share: the use of underspecified antecedents. In order to address both approaches, it is helpful to have an example that highlights where they diverge from each other. Suppose a bomb capable of very quickly destroying Earth is set to activate when a neutrino interacts with the trigger. Neutrinos are ubiquitous in nature but each only rarely interacts with ordinary matter. In actuality, the bomb never activates even though all it takes for activation is for a single neutrino to be ever so slightly in a different location. Now, consider an ordinary counterfactual having nothing to do with the bomb. Let us say that Guy completed an ordinary workday at the office with his ever-observant boss. By ordinary standards, the following is true: If Guy had not shown up for work, his boss would have noticed. According to the informal approach, we are free to evaluate this counterfactual by examining situations where Guy spends the day goofing off or is sick or sets out for a new life of adventure or dies or some disjunction of these and other plausible ways Guy could fail to show up for work. We are also free to include the possibility where the reason Guy does not show up for work is that the Earth has exploded, but we are not required to. For this particular example, most reasonable ways of filling out the antecedent state lead to the counterfactual being reasonably highly probable, or loosely speaking, true. (If one imposes the universal modality principle, it will

[8] If there is any fundamental chanciness, one conditionalizes on the truth of the antecedent to get the appropriate probability measure over future evolutions.

turn out false, because the boss would not necessarily notice Guy's absence, but on such a reading, virtually all mundane counterfactuals turn out false.) By the principled approach, one needs to consult one's theory of counterfactuals. If that theory identifies the bombed-Earth worlds as the theoretically specified worlds relevant to the semantic value of the counterfactual, then the counterfactual will be definitively false (or highly improbable) because the boss will not exist. This illustrates that the two approaches can differ on how to evaluate the same counterfactual statement.

The informal approach is good as far as it goes. However, its capacity for optimizing our understanding of effective strategies is non-existent because it basically collapses into the map approach, and thus its applicability to influence is parasitic on counterfactuals that use only sufficiently specified antecedents. As the bucket example illustrates, our natural attitude toward mundane counterfactuals is to treat many of the physical background conditions not as part of the stipulated content of a more thoroughly fleshed out antecedent but as part of the objective facts that play a substantive role in setting the semantic value of the counterfactual. But if we are using the informal approach, we are in effect *choosing* the antecedent event rather than deriving it from some theory of counterfactuals. Though we might be guided by some rules of thumb governing what the appropriate antecedent event should be, if our final judgment about how to characterize the antecedent event just comes from a free choice of some modification to make the antecedent obtain, then we are effectively stipulating an antecedent event for any given actual world. That is the map method.

The main deficiency of the principled approach is that it imposes a restriction on the set of counterfactual worlds that is suboptimal for clarifying the nature of causation. In the case of a pure forking theory, this exhibits itself in several ways. When the antecedent event, C, cannot be brought about through an indeterministic evolution, the forking theory can at best treat it as a counterlegal, a counterfactual whose antecedent is nomologically impossible. For example, if conservation of mass holds as a fundamental law, then a counterfactual postulation of the form, "If everything at time t were the same as in the actual world except for an additional massive corpuscle at location p, then ...," cannot be addressed by the forking account in an informative way because the antecedent event is counterfactually impossible.

When the antecedent event, C, can be brought about through an indeterministic evolution, there are other problems. Pure forking accounts can be subdivided into those that provide a rule for when the forking time, t_f, occurs and those that treat t_f as an additional input parameter. Theories that dictate the appropriate forking time attempt to do so as a means of keeping the past counterfactually fixed as much as possible without requiring a bizarre evolution to make the antecedent obtain. For example, suppose the fundamental laws are those of spontaneous collapse interpretations of quantum mechanics and consider a patch of land where there are no fossil-like objects. (A fossil-like object is an object that is physically just like a fossil, but it does not need to have been created from a process of fossilization.) The counterfactual possibility where there is a fossil-like object in that patch of land could be interpreted as the result of a fork immediately before the present which generates a highly improbable quantum collapse leading to the spontaneous generation of a fossil-like object, a bizarre evolution. Or it could be interpreted as the result of a fork hundreds of millions of years ago generating a fossil in ordinary ways. Forking theories that dictate a forking time must choose some time within that span. To the extent they select more recent times, they artificially make it impossible for the fossil-like object to be the fossil of earlier species of dinosaurs. In the extreme case where the forking time is as recent as possible, that forces the fossil-like object to appear spontaneously. To the extent forking theories make the forking time further in the past, they increase the amount of counterfactual alteration to the present, however much context dictates that the present stays fixed. For example, the laws might dictate that the kind of changes in the distant past that are needed to make a fossil exist in the present patch of land are enough to make it unlikely that contemporary society exists.

Forking theories that leave the t_f as a free parameter are superior because they provide more freedom to accommodate the intended counterfactual possibility. However, they still impose an unnecessary restriction on how the antecedent comes about. For example, even if one's interest in Guy has nothing to do with the bomb, any t_f one selects will force a portion of the possible worlds to be those where the Earth is destroyed. If the point of counterfactual dependence were to pronounce on 'what would have happened in a singular instance of history had things been different', then it would be reasonable that the particular character of the actual world at time t_f should play a role in fixing which counterfactual possibilities are relevant to the semantic value of the conditional. However, if we are interested in what follows generally from the intended hypothesis that Guy failed to show up due to some ordinary reason, we would be unable to exclude the worlds where the Earth is destroyed because the original proposition did not exclude them. Of course, in practice, someone who was not interested in including the bombed-Earth worlds and realized that those worlds were being counted as relevant could just reformulate the intended counterfactual to exclude them. But such a maneuver is an admission that the principled approach is insufficient for its intended purpose.

The other known example of the principled approach is Lewis's (42) theory of world-similarity. The two core ideas behind Lewis's theory of counterfactuals (40)(41)(42) is that the semantics governing counterfactuals is based on truth conditions incorporating a relation of comparative similarity and that a principled theory of world-similarity can instruct us of the truth of counterfactuals pertaining to nomic interactions among worldly stuff. The semantics itself is simple enough. The antecedent and consequent are modeled as propositions. A counterfactual $C \ \Box\!\!\rightarrow E$ is true if and only if C is a contradiction or for some $(C\&E)$-world, there are no possible $(C\&\neg E)$-worlds that are more similar to actuality than it. E

counterfactually depends on C if and only if $(C \mathrel{\Box\!\!\rightarrow} E)\&(\neg C \mathrel{\Box\!\!\rightarrow} \neg E)$. Even though the criticisms in the previous subsections suffice to show that Lewis's logic and semantics are suboptimal for understanding influence in an empirical analysis, Lewis's substantive theory of how to address underspecified antecedents is still worth some investigation. I will not address any of the alleged counterexamples to Lewis's theory of world-similarity but will merely point out a single problem with his approach that will likely extend to any other principled approach to the problem of underspecified antecedents. The failure of both the informal and principled approach then casts doubt on the utility of any theory that employs underspecified antecedents, including all approaches to counterfactuals based on a similarity relation.

Lewis's (42) theory of world-similarity provides a partial function whereby one can input the actual world and two other worlds, w_1 and w_2, and receive as output an identification of which world is closer to actuality according to the similarity relation. The theory does not pretend that its pronouncements on the truth of some counterfactual will match common sense judgments or even philosophically refined judgments, but instead is intended to be restricted to the so-called "standard resolution," which is applicable to counterfactuals where the antecedent and consequent pertain to the occurrence or non-occurrence of events. The standard resolution is not defined neutrally in terms of patterns of linguistic data concerning which counterfactuals people standardly agree to but is a theoretical contrivance designed to bolster Lewis's (41) account of causation. The important feature of Lewis's theory of similarity for our purposes here is that it compares possible worlds using only features that are relatively simple in physical terms. Except for some fudge factors baked into the theory, it assesses world-similarity only on the basis of the spatial extent to which the evolution of the material content in a given world counts as a violation of the laws of the actual world and the time span during which that world perfectly matches the material contents of the actual world. To the extent it is based on a relatively simple system of relatively simple, non-anthropocentrically framed parameters, it deserves to be identified as a principled theory of counterfactuals.

It is easy enough to see why an appeal to some sort of similarity is useful for handling underspecified antecedents. The worlds we are interested in for assessing influence are those that result from the lawful evolution of states that are just like the actual world except adjusted to make the antecedent obtain. So we need some structure that gets us from a proposition and an actual state to the counterfactual state we want to consider. Since we also have a defeasible rule in our linguistic pragmatics that tells us not to modify the actual state more than is appropriate to instantiate the truth of the antecedent, we already have something like a command to find the worlds where the antecedent is true that are most similar to actuality.

However, in order to be optimal for understanding the empirical phenomena motivating our notions of influence, it needs to come close to matching the results of the minimal account, but this is going to be difficult to accomplish because of the problem of negative antecedents. For Lewis, counterfactual dependence between two existing events, C and E, tracks the truth value of $\neg C \;\Box\!\!\rightarrow\; \neg E$. But when C is a mundane event, the proposition $\neg C$ has a vast extension that includes all sorts of possibilities that are intuitively irrelevant to the truth of the counterfactual. The point of the similarity relation itself and the principled theory of world-similarity is to narrow the $\neg C$-worlds to a respectable set that near enough corresponds to the possibilities identified by the minimal account. The problem of negative antecedents is that many negations of mundane events are intuitively irrelevant to the intended counterfactual, and they are very hard to exclude from the set of nearest possible worlds without violating the spirit of the minimal account. For example, in Lewis's own specific proposal, the closest world to actuality where the proposition 'Guy does not attend work' holds includes worlds where the bomb activates and destroys the Earth.

What is intuitively wrong with that result is that the point of reasoning about what would have happened had Guy not shown up for work is to explicate some reliable inferences from the general condition of Guy somehow not showing up to work. We intuitively know more or less the right possibilities to consider, and evaluating counterfactuals via a theory of world-similarity opens up the possibility that the scenarios we think are relevant turn out to be entirely irrelevant. The mere fact that we can be wrong about which possibilities are relevant to the correctness of some counterfactual is not a problem in itself. What is troubling is that the truth value of the counterfactual will be unhelpfully sensitive to nuances in how the antecedent is characterized when clearly such nuances have no useful role in understanding influence. For example, Lewis's theory dictates that it is false that if Guy had not shown up for work, his boss would have noticed, but true that if Guy had done something other than show up for work, his boss would have noticed. Even more troubling is that when there is disagreement between the worlds picked out by theory of world-similarity and those by the plain intent of the explicitly stated antecedent, using the theory of world-similarity provides a less reliable guide to tests of influence. To test the intended 'Guy' counterfactual, we ought to have Guy stay home from work and then try to watch his boss, not destroy the Earth and then try to watch his boss, moral considerations aside.

The important conclusion is not that Lewis's own particular scheme for evaluating world-similarity fails to capture the relevant worlds, but that any theory in its neighborhood will also fail to address the problem of negative antecedents. Because of the predominance of bizarre worlds, there are always plenty of ways for the universe to fork so as to bring about the non-existence of some mundane event. It does no good to rule them out merely on grounds of probability because in the deterministic case, the forking comes about through miracles and there is no probabilistic constraint on forking miracles. It does no good to dismiss them as artifacts of a very unusual situation, for such Earth-destroying possibilities are dynamically possible (though extraordinarily improbable) according to any indeterministic interpretation of quantum mechanics. Also, as shown by Elga aE2001, the actual world has ubiquitous time-reverse equivalents of the neutrino-triggered bomb. Rejecting the problematic bizarre worlds merely on grounds of their bizarreness threatens to bias the counterfactual outcomes toward the kind of influence relations humans conceive of as natural. Most important, the bizarre worlds cannot be brushed off as contextually irrelevant because the whole point of the similarity relation is to narrow the space of possibilities where the antecedent is true to just those that are relevant to the truth of the counterfactual. To insist that one should just reformulate the antecedent to rule out the possibility that the bomb is triggered is just to give up on a principled account of underspecified antecedents. It is no more a justifiable maneuver than Jane's after-the-fact attempt to rig her counterfactual with additional constraints to make a water-filled bucket more likely.

To summarize, there are two methods of handling antecedents as underspecified propositions, both of which fail. The principled approach suffers from an inability to pick out the relevant worlds and the informal approach only succeeds to the extent that it cheats by treating antecedents as sufficiently specified. Thus, treating antecedents as underspecified propositions is inferior to treating them as sufficiently specified. Because the nomic conditional treats antecedents as sufficiently specified, it is better suited for an empirical analysis of causation focused on effective strategies.

{ 12 }

Morgenbesser's Coin

In this chapter, I will further illustrate the character of the nomic conditional using the example of Morgenbesser's coin to highlight the most important differences between the nomic conditional and natural-language-based models of counterfactuals. The goal of this chapter is to inoculate the account of general causation presented in chapter 5 from some misinterpretations that can occur if readers are not fully alert to the significant respects in which prob-dependence differs from versions of counterfactual dependence based on standard models of counterfactual conditionals. In the chapter on causal asymmetry, I contrast the nomic conditional with Barry Loewer's (45) SM-conditional based on statistical-mechanical probabilities.

The example known as Morgenbesser's coin involves a bet made on a standard coin flip where we assume the existence of stochastic dynamical laws with enough microscopic randomness to ensure that the chanciness of an ordinary coin flip is overwhelmingly the result of fundamental stochasticity rather than the coin's precise initial conditions. The scenario begins with Jane betting by calling 'heads' as Jill flips the coin, and the coin lands tails. The Morgenbesser counterfactual is, "If Jane had bet tails, she would have won." A plausible explanation of why people assent to the Morgenbesser counterfactual goes as follows. We have a default psychological heuristic for evaluating counterfactuals that starts by modifying the actual history of the world in order to make the antecedent true, and then tracing any consequences of that alteration using our hypotheses about how the world operates. We imagine the situation as it was just before Jane decided to call 'heads.' We then imagine that chancy events in Jane's brain lead to her deciding to call 'tails.' We trace the evolution of this scenario into the future by reasoning that Jane's calling 'tails' will not affect the coin flip because in all normal circumstances voicing one word rather than another does not affect the probabilities of coin flip outcomes. Because the actual outcome was tails and no differences in the counterfactual scenario motivate us to reassess the actual outcome, we reason that the coin flip in the counterfactual scenario landed tails and thus that Jane would have won the bet.

To investigate the Morgenbesser counterfactual, I will now explore several ways to make this somewhat speculative hypothesis about how we evaluate counterfactuals more rigorous and more general. The resulting procedure will not capture our implicit reasoning for all counterfactuals, but this is not a serious limitation because it is only intended to do well enough (when the counterfactuals involve worldly happenings) to clarify a range of approaches to the semantics of counterfactual statements.

According to the sketched procedure, to evaluate counterfactuals of the form, "If C had occurred (at time t),...," one should imagine a state S at t that instantiates C. This could be done by modifying the actual state of the world at t (if necessary) to instantiate C. Or S could be constructed by starting with the actual state at some suitable time before t, evolving this state forward to t under the laws of nature, and then conditionalizing on the existence of C. Then, further reasoning concerning the evolution of S would be employed to ascertain whether E obtains. These two ways of carrying out the initial stage of the inferential procedure can then be supplemented with a rule for how to propagate the counterfactual state forward in time. Here are some possible options:

1. ***Change Nothing***: The simplest method of evolving S forward in time is just to fill in the future of S with exactly what happens in the actual world. This method is virtually never employed when the antecedent is false—for obvious reasons—but when the antecedent is true, it is standard practice to let S be the actual fine-grained state at t, even when the actual state does not instantiate something that ought to count as among the situations contextually relevant to the evaluation of the counterfactual. Then, instead of using the laws to evolve S forward, we just fill out its future with the actual future. The 'change nothing' procedure has the consequence that when C is true, the truth value of "If C had happened, E would have happened" matches the truth value of E. This makes sense of several theorems of counterfactual logic, especially $(C\&E) \vdash (C > E)$ and $(C\&\neg E) \vdash \neg(C > E)$, though I am not taking sides on whether these are good principles of reasoning.
2. ***Change Everything***: Another simple method for evolving S forward in time is just to let the future be whatever S fixes for the future. This method corresponds with the following way of reasoning about contrary-to-fact possibilities: One imagines hypothetically going back to some time, fiddling with the state if necessary to make the antecedent true in some contextually relevant way, and then let the laws dictate what happens afterward without any bias due to (later) accidental contingencies in the actual world. One is figuratively rerolling the dice of nature for the counterfactual history starting with S, as depicted in Fig. 12.1. Following this method, we would say that if Jill had bet tails, there would have been a fifty percent chance of her winning, but no fact of the matter as to which outcome would have happened.[9]

[9] Philosophers tend to emphasize that natural language counterfactuals have a universal-quantifier character. When the 'change everything' procedure leads to no definite outcome for Jane, we say it is false that she would have won and false that she would have lost. The implicit

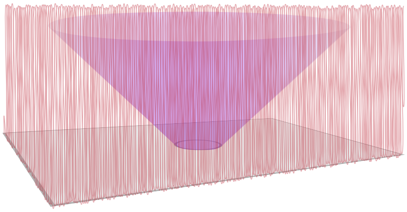

FIGURE 12.1 *Change Everything: Reroll the dice of nature for everything after t regardless of whether the antecedent is true.*

3. ***Bifurcated***: The bifurcated method is simply to apply the 'Change Nothing' method if the antecedent (construed as a proposition) is true, and to apply the 'Change Everything' method if it is false. This is motivated by the observation that the 'Change Nothing' method seems to match our intuitions well when the antecedent is true, but greatly mismatches our intuitions when the antecedent is false.
4. ***Change Infected Regions***: It is possible to match people's instinctive judgments of particular counterfactuals better than the previous options by evolving the antecedent state forward in time while selectively employing either the actual world as a guide or the dynamical laws. We can do so by drawing a distinction between an infected and uninfected region of spacetime.[10] The way it works is as follows. We start with the actual state, $S_@$, at time t and check to see whether the actual world around $S_@$ instantiates C. If it does, we declare our initial hypothetical state S to be $S_@$ and we declare that it counts as entirely uninfected. If $S_@$ does not instantiate C, we set S to be just like $S_@$ except that we modify it to make it instantiate C in some contextually appropriate way. One can typically accomplish the alteration by just modifying a local patch of physics and leaving everything else exactly the same, or one can back up in time and let the laws evolve to t and then conditionalize on C. Whatever patch is modified to instantiate C counts as infected, and everything else counts as uninfected. Now, we evolve S forward in time to create a counterfactual scenario, a family of counterfactual worlds that count as the worlds relevant to the truth of $C > E$. The two rules for evolving the state forward in time are these: Wherever S is uninfected, we just copy whatever happened in the actual world into the counterfactual scenario. Wherever S is infected, we use the actual dynamical laws to

assumption is that "If Jane had bet tails, she would have won," should be interpreted as "...would definitely have won." Furthermore, philosophers almost always think of this interpretation as part of the semantics of the counterfactual conditional and not some pragmatic factor. However, one does not need to treat counterfactuals in this manner. In fact, in ordinary usage, people often assent to counterfactuals even when knowing that there is some chance that the consequent will not follow from the antecedent, which might be interpreted as evidence that in at least some cases, people think of counterfactuals probabilistically (in degrees) rather than as having a binary truth value. Adherents of maintaining some sort of modal character in the counterfactual semantics rather than the pragmatics are partially able to address these cases using such devices as similarity relations or context fixing parameters to maintain the truth-based semantics.

[10] I am lifting this distinction from Maudlin (48), p. 30.

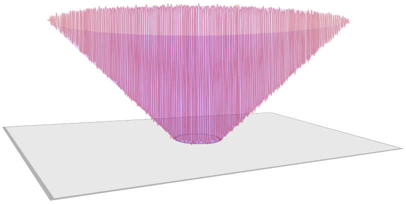

FIGURE 12.2 *Infection by Contribution: If C is true, change nothing. If C is false, reroll the dice of nature for everything inside the future light cone of the antecedent event C.*

fill in the counterfactual scenario by evolving the material contents of the counterfactual scenario just prior to the occurrence of the infected region and propagate them lawfully into the infected regions. There are different prescriptions for the rules as to when a region of space-time counts as being infected. Let us now examine some possible rules for infection in order to get a better feel for the methodology. No matter which rules we use for how infection spreads, the resulting system satisfies the three previously mentioned principles common to counterfactuals based on natural language: centering, actuality-focusing, and antecedents as underspecified propositions.

- Infection as Contribution: One result of the 'change everything' method that sounds intuitively incorrect is that it does not respect the principle that anything entirely causally disconnected from the counterfactual alteration should be unaffected by the alteration. Assume that relativistic locality holds—i.e. that the arena is Minkowski space-time and that non-spatiality[11] holds—and that the entire coin flip process happens at space-like separation from Jane's counterfactual bet on tails. Intuitively, because the bet is then not a causal contributor to the flip outcome, hypothetically altering the bet ought to make no difference to the outcome, and so the actual outcome should be held fixed as the counterfactual outcome. The method of 'infection as contribution' extracted from Maudlin (48), p. 30 says that any infected region of the arena infects its domain of influence, as depicted in Fig. 12.2. In the case of the Morgenbesser coin, that means everything outside the light cone of Jane's calling 'tails' is held fixed, i.e. kept just like it is in actuality, and everything in the future light cone of Jane's call is recalculated using the fundamental laws from the boundary conditions on the light cone. This provides a probability for all events including whether Jane wins the bet.
- Infection as Probability-Affecting: A more restricted conception of infection (37) (38) (50) (51) is that one should count as infected any event whose occurrence is probabilistically dependent on events from

[11] Remember that non-spatiality is the principle that events never termine events that are at space-like separation.

infected areas. Because the chance of the coin landing tails is arguably very nearly the same whether Jane bets heads or tails, it is probabilistically independent of the bet and so is deemed uninfected. Because Jane's winning the bet does probabilistically depend on her bet when one holds fixed the actual flip outcome, it counts as infected. Thus, by using the laws to fill in the infected region, we get the result that the counterfactual comes out true. There are serious questions about how to make the notion of probabilistic dependence suitably precise without employing counterfactuals that are circular or relying on parameters that would leave the semantic value of the counterfactual too subjective, but I will leave these issues aside.

- Infection as Culpable Cause: Another variation of this idea, (16) (59) p. 306–7, is to count as infected any location where events "causally depend" on what happens in infected regions. What counts as causal dependence for the purposes of determining infection presumably relies on intuitions about causal culpability. Both Edgington's and Schaffer's evaluations of Morgenbesser's coin count the coin outcome as not causally dependent on the bet, where 'causally dependent' is intended to invoke a less inclusive notion of causation than contribution. I suppose the idea behind their views is that there are culpable causation facts that amount to something more than just handy talk for contribution and probability-affecting, and further that there are generalities regarding not only actual culpable causation but possible culpable causation. Perhaps the implicit reasoning is that in virtually all Morgenbesser cases, the bet is not a culpable cause of the flip outcome. From this general pattern of no culpable causation, we infer that calling out a bet one way rather than another is in general not a culpable cause of coins landing one way rather than another. Thus, we should not count the flip outcome region of space-time infected by virtue of the infected bet-calling region. Schaffer notes that a more sophisticated understanding of the relevant causal notion could take into account various microscopic ways in which the bet affects the outcome. The semantic value of the Morgenbesser counterfactual would then depend on the particular construal of 'causally dependent.' If one employs a notion of causal dependence thoroughly stripped of bias from pragmatic simplifications, one gets a resolution resembling 'Infection by Contribution,' which makes it either false or fifty-percent probable.[12] The intended purpose of 'Infection as Culpable Cause' is to make the Morgenbesser counterfactual come out true, so presumably a more restricted notion of causal dependence is intended.

[12] Which option is appropriate depends on whether one construes the counterfactual as having a universal-quantifier character as discussed in the footnote at the end of the 'Change Everything' option above.

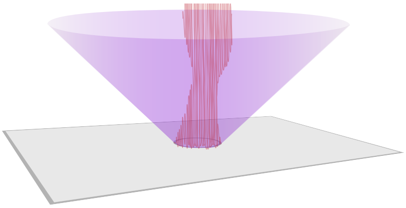

FIGURE 12.3 *Infection by Culpability: If C is true, change nothing. If C is false, reroll the dice of nature for everything that causally depends on C in some folksy sense of 'cause'.*

It is frequently unclear whether the models listed above are intended by their advocates as psychological models for explaining people's intuitive judgments about the correctness of counterfactual claims, or whether they are intended as part of a metaphysical model with consequences for influence and causation. Without a clear enough guide to their purpose, there is no way to decide which methods for assessing the Morgenbesser counterfactual are better than others.

Given that my aim is to provide an empirical analysis of the metaphysics of causation, it is important for my nomic conditional to be oriented metaphysically without making any effort toward vindicating the explicit truth of commonsense opinions. Insofar as one is concerned with metaphysics, the alleged truth of the Morgenbesser counterfactual is not a datum that needs to be accounted for. We can have different attitudes toward the "obviously correct" Morgenbesser counterfactual without these intuitions making any difference to any empirical phenomena that reveal how things in the world operate.

Insofar as one is concerned with our folk theory of influence, our inclination to agree with the Morgenbesser counterfactual is a datum that needs to be accounted for. So it is important for a comprehensive empirical analysis of causation in the special sciences to be able to explain why people typically find the Morgenbesser counterfactual agreeable. The important methodological point is that we can account for this psychological datum by arguing that our inclination to agree with the Morgenbesser counterfactual results from our employing a patchwork of psychological heuristics that may not ultimately cohere with one another in a complete, precise system. One does not need to explain this empirical phenomenon using a model of metaphysics that renders the Morgenbesser counterfactual explicitly true.

To see how our instinctive judgment concerning the Morgenbesser counterfactual is irrelevant not only to the explanation of effective strategies but to any empirical analysis of the metaphysics of causation, it is enough to conjoin the following observations.

First, all remotely reasonable methods of evaluating the Morgenbesser counterfactual presuppose two kinds of facts: those that concern the material content of the actual world in the region *R* where the Morgenbesser scenario takes place and some sort of laws that are applicable to actuality as well as to any relevant contrary-to-fact situations. The invoked laws could be understood liberally to include rules of thumb or even miracles.

Second, the material content of *R* and the operative laws are uncontroversially empirically accessible in the intended sense. Although there may be limits on the extent to which we can gather information about the microscopic positions of things and the extent to which we can get an accurate grasp of the laws of nature, it is well accepted that we can empirically test non-trivial hypotheses about them.

Third, there is nothing about the Morgenbesser counterfactual itself that is testable within the region *R* beyond what is already testable about the material content of *R*. One cannot empirically test claims about what would have happened in *R* if things had happened other than the way they happened.

Fourth, there is nothing about the Morgenbesser counterfactual itself that is testable in regions other than *R* beyond what is already testable about the actual laws.[13] What would have happened in *R* had Jane bet tails does not carry over to any new situation that starts the same way as *R*. For example, if you flip the same coin again in exactly the same kind of physical situation and have Jane bet tails, the operative laws will govern what happens without any dependence on what would have happened counterfactually in *R*. The chance outcomes of one situation make no difference in how the laws operate elsewhere.[14]

The empirical phenomena motivating talk of counterfactuals do not go beyond (1) what we can get from the laws (including various global contingencies like the values of fundamental constants) when we are concerned with testing claims outside of *R* and (2) what we can get from the particular material content of *R*. Thus, the only way to privilege one interpretation of how to evaluate the Morgenbesser counterfactual over others is merely by virtue of how it optimizes explanations of the general laws and the material content in *R*.

What makes the Morgenbesser counterfactual irrelevant to metaphysics is that its evaluation presupposes two kinds of facts—facts about the material contents in *R* and facts about the laws—such that once we have optimized those toward empirical phenomena, any additional factors one incorporates to make the Morgenbesser counterfactual turn out true will have the side effect of making the account of counterfactuals less optimal for assessing general tendencies or causal generalizations. (The additional factors here include anything in one's account of counterfactuals that restricts the infected region more severely than 'infection by contribution'. For example, the method of 'infection as probability-affecting' and

[13] One should also include general facts about the material content such as the topology of the arena, values for fundamental constants, and the kinds of fundamental interactions, even if they are not bona fide laws.

[14] In making this claim, I am not assuming that the fundamental laws disallow that what happens in one location causally contributes to what happens elsewhere. Quantum mechanical entanglement, for example, might ensure that everything in the universe is linked in a such a way that a counterfactual alteration to one location determines differences everywhere else, but the scenarios being entertained in my discussion ex hypothesi do not fundamentally interact with one another in any interesting way.

the method of 'infection as culpable causation' both attempt to hold certain aspects of the actual future fixed under the counterfactual alteration.) In order to get the Morgenbesser counterfactual to come out true, the additional factors—no matter what else they include—must contain the actual outcomes resulting from fundamentally chancy processes. But it is exactly these actual outcomes that we do not want to include in our predictions of what happens in regions other than R because it is simply the nature of distinct fundamental chance outcomes (that are independent of one another insofar as fundamental causation is concerned) that what happens by chance in one case does not have any bearing on what happens by chance in other cases. So, the machinery needed to make the Morgenbesser counterfactual true works against our having an optimal guide for predicting and explaining what happens elsewhere.

In arguing that the explicit truth of the Morgenbesser counterfactual is irrelevant to metaphysics, I am not arguing that in order for a concept X to be relevant to metaphysics it must be empirically testable against rival candidate concepts Y, Z, etc. After all, my own conception of influence as prob-influence cannot be empirically tested to see whether it is better than influence as contribution. What makes prob-influence a better conception of influence than contribution is that it serves better in an overall account of the empirical phenomena associated with influence and causation. Its superiority is patently not an empirical issue. What makes a concept *metaphysically* valuable is that *what it is aimed at optimizing* is empirical. The methods designed to make the Morgenbesser counterfactual come out true, e.g. 'Infection as Probability-Affecting' and 'Infection as Culpable Cause', do not optimize the counterfactual conditional in ways that help it to apply to what actually happens in R or outside R. Thus, it is optimized to fit something other than empirical facts,[15] and that makes it suboptimal for metaphysics given that we already have law facts and material facts in the metaphysical system.

Before concluding this chapter, I will make a final clarification of the relative merits of the various methods of evaluating counterfactuals. Recall that my method of counterfactual evaluation is to settle on the antecedent event by just stipulating a contextualized event $\overline{C}$. If one wants, one can arrive at $\overline{C}$ by setting it to be that which one gets by minimally modifying the actual state at t to instantiate the truth of the antecedent proposition C, but such a procedure is not required. On my account, one has complete freedom to start with whatever contextualized event one wants. By contrast, all the above methods use some sort of minimal modification. This means that they are suboptimal by virtue of obeying actuality-focusing and may also be suboptimal by virtue of using a principled approach for precisifying antecedents that are initially underspecified propositions. All methods except 'change everything' are also suboptimal because they obey centering.

[15] Again, I am setting aside empirical phenomena relevant to the evaluation of our theories of how people think about counterfactuals.

We could overcome these previously discussed problems by revising the methods listed above to be far more liberal about what counts as a minimal modification to instantiate the antecedent. Specifically, let us now replace the step in the above methods that said, 'alter the actual state minimally to make the proposition C true' to 'instantiate $\overline{C}$ in place of the actual state'. When we do, we obviate problems with centering, actuality-focusing and antecedents as underspecified propositions.

With this modification in place, it turns out not to matter too much whether we employ 'change everything' or 'infection by contribution' in my method for evaluating prob-dependence. So long as $\tilde{C} \equiv (\overline{C}, \overline{\neg C})$ occupies a portion of a single time slice, we will only get prob-dependence of events on $\tilde{C}$ when they are in the domain of influence[16] of the region where $\overline{C}$ and $\overline{\neg C}$ disagree about the material facts. For example, suppose we have a naturally contextualized event, $\overline{C}$, involving a match being struck in region R and a contrast $\overline{\neg C}$ that is exactly the same as $\overline{C}$ except for having a prototypical lack of a match strike in region R. Furthermore, suppose the laws are relativistic and consider some consequent E that is a coarse-graining of a fundamentally chancy event e that actually occurred outside the light cone of R and to the future of $\tilde{C}$. According to my method, we evaluate both counterfactuals, $\overline{C} \;\square\!\!\Rightarrow\; E$ and $\overline{\neg C} \;\square\!\!\Rightarrow\; E$, by rerolling the dice of nature throughout the future of $\overline{C}$ and $\overline{\neg C}$, ignoring that E actually occurred. Although both counterfactuals will have a non-trivial value (equal to the chance given either $\overline{C}$ or $\overline{\neg C}$), their values will be exactly the same, which implies no prob-dependence.[17]

If we were to replace my method by only rerolling the dice of nature for the future light cone of R, we would get the same result. In that case, both counterfactuals would have value 1 because E was an event that occurred in every $\overline{C}$-world and every $\overline{\neg C}$-world. So again, there would be no prob-dependence; hence, no prob-influence.

If we were to evaluate counterfactuals with different regions of modification, R and R', then we would get misleading results whenever there is enough fundamental stochasticity because there would be regions where the domain of influence for R would not coincide with the domain of influence for R'. In such regions, one counterfactual would in effect reroll the dice of nature and the other would use the actual material contents. This would lead to misleading proclamations of prob-influence.

In light of the fact that the modified method of 'infection by contribution' gives the same result as mine when the regions are the same, but different results when the regions are different, it is best just to stick to the method as I presented it earlier and drop the whole idea of modifying it to respect intuitions that 'infection by contribution' is the correct way to evaluate counterfactual claims.

[16] Recall the definition of a domain of influence from §2.6.

[17] This equality holds because e's domain of contributors, e's past light cone, intersects with both $\overline{C}$ and $\overline{\neg C}$ in the same region including exactly the same material content. (Recall the definition of a 'domain of contributors' from §2.6.)

The methods 'Infection as Probability-Affecting' and 'Infection as Culpable Cause' do not fare well even with the helpful modification in place. What is most problematic about them is that they provide insufficient precision. There is no guide as to how much probability-affecting counts as enough to trigger infection and because any threshold will be arbitrary, some clarificatory stipulation will need to be added. The same goes for clarifying the precise extent of the supposed relations of causal dependence governed by intuitions about culpability. With any such threshold parameters, there will be cases where a large difference in the value of the counterfactual will occur because of a small change in the standards for what triggers the spread of infection. This results in ungraceful conceptual degradation.[18] Furthermore, the parameters to make either method precise have been notoriously hard to pin down. I think all such avenues of exploration amount to attempts to explain the clear in terms of the murky.

One might wonder why we tend to share the judgment that the Morgenbesser counterfactual is true, given that its truth is irrelevant to anything empirical. I will just make a brief observation here. The answer, I think, is that it is an imperfect patch for the problems created by having an implicit conception of counterfactual dependence like that of the bifurcated notion of influence discussed above. On the one hand, it is understandable that we would want counterfactual conditionals to obey modus ponens. Furthermore, although it is doubtful that we employ centering as a general principle of counterfactual reasoning, it is plausible that we sometimes implicitly use centering when making retrospective judgments about the counterfactual dependence among past events by comparing the unique actual world with a range of possible counterfactual worlds. On the other hand, it is also understandable that we think of contrary-to-fact conditionals concerning worldly happenings in probabilistic terms, even if we cloak this probability with semantic devices to make it truth-apt. That is, we implicitly recognize that contrary-to-fact scenarios are compatible with a range of different outcomes without any one of them being special in the way that the actual world is special. Furthermore, we recognize that it is often useful to think of some counterfactual outcomes as more probable than others. When we try to maintain both a special role for the actual world and a probabilistic treatment of the contrary-to-fact worlds, the result is something in the neighborhood of the bifurcated notion of influence. The problem with the bifurcated notion, recall, was that it led to misleading implications regarding influence because it is unable to distinguish between (1) counterfactual dependence that arises merely because E was a freak accident and (2) counterfactual dependence that arises because laws of nature relate C to E. Because we often use counterfactuals to express causal culpability (and indirectly nomological dependence of a probability-affecting kind), uttering a counterfactual that implies (via the bifurcated notion) that E counterfactually depended on C will often convey that C influenced E. Because that implication is misleading when

[18] See §1.1.

the counterfactual dependence only arose because *E* was unlikely, it is helpful to have a notion of counterfactual dependence that mitigates the undesired implication. By carrying over from the actual world to the counterfactual worlds any chance outcomes that would result in misleading counterfactual dependence, one automatically eliminates much of the spurious dependence. But because we need to allow that some chance outcomes are not held counterfactually fixed, we need a principle that rules out spurious dependence while also ruling in genuine dependence. Accomplishing that task, I take it, is the purpose of 'Infection as Probability-Affecting' and 'Infection as Culpable Cause'. Because such construals of counterfactual dependence are psychological kludges, it is not surprising that 'Infection as Probability-Affecting' and 'Infection as Culpable Cause' are unimpressive at clarifying the nature of causation and influence.

{ 13 }

Orthodox Conceptual Analysis

This chapter was excluded from the printed version of *Causation and Its Basis in Fundamental Physics* because there is at present widespread hostility toward empirical analysis. I offer my thoughts below only for the possible benefit it may provide graduate students. More advanced readers should skip this chapter.

Because my methodology for approaching the metaphysics of causation is unfamiliar, in this chapter I will contrast my version of empirical analysis with what can be called '***orthodox conceptual analysis***', or just 'orthodox analysis' for short. I will also offer some arguments for adopting empirical analysis over orthodox analysis.

Any conceptual analysis of X can be thought of as a systematization of the platitudes that constitute our implicit concept of X. To conduct a conceptual analysis, we start off with some initial data in the form of uncontroversial truths about the concept, including paradigm examples of the concept, known as ***exemplars***, as well as a priori links to other concepts. For example, an orthodox analysis of food would begin with propositions that an orange is food, a hoagie is food, etc., as well as with broader principles that food is the kind of thing people typically like to eat, the kind of thing that relieves hunger, a kind of material substance, a category that is species-relative, etc. These naive platitudes are then systematized in some principled way. The principles that govern the standards of adequacy for orthodox analysis vary quite a bit among those who practice it. I will first survey a range of prominent opinions from advocates of the orthodoxy and then return to summarize some necessary conditions accepted by all versions of the doctrine.

What I call '***old-fashioned orthodox analysis***' attempts to systematize toward an explicit definition, a statement of the form "x is food if and only if . . .," where the dots are to be filled in with necessary and sufficient conditions in terms of concepts that are distinct enough from food to avoid conceptual circularity and are principled enough. Ducasse's (14) discussion is a good example of an old-fashioned orthodox analysis of causation. For an analysis to be principled enough is for its explicit definition to avoid being merely data fitting. It is unacceptable, for example, to list a bunch of exemplars of food and exemplars of non-foods and then claim as one's definition that any substance that is sufficiently like the food exemplars and sufficiently unlike the non-food exemplars is a food. Without further specification of what 'sufficiently like' amounts to, the analysis is nothing more than a summary of common sense intuitions. This deficiency cannot be remedied merely by collecting survey data about which substances people think of as food and then specifying a mathematical function that best fits the data. Such a scheme fails as a conceptual analysis because it provides no interesting account of what foods have in common that non-foods fail to share.

The hallmark of old-fashioned orthodox analysis is an insistence on a theory's matching strong folk intuitions. David Lewis (43) adopts this standpoint by declaring that "[w]hen common sense delivers a firm and uncontroversial answer about a not-too-far-fetched case, theory had better agree. If an analysis of causation does not deliver the common-sense answer, that is bad trouble." Also, "when our opinions are clear, it's incumbent on an analysis of causation to get them right." (44) What it means to "get an opinion right" or "deliver the common-sense answer" is insufficiently clear, I think, but the rough idea is that the kind of match demanded cannot be mediated through an account of how the common-sense opinion is a false but understandable simplification of reality. In this chapter, I will attempt to explicate this guiding idea as much as possible, but in order to do so, it helps to have a label for the demanded connection. So, let us say that a conceptual analysis renders a statement S ***explicitly true*** when the analysis declares that S is true in the most straight-forward literal sense rather than declaring that S is strictly speaking false but understandable as true in light of practical concerns. For illustration, consider a conceptual analysis of food holding that a substance is food if and only if it is nutritious. Under such a conceptual analysis the claim, "A hoagie is food," is rendered explicitly true because our (correct) common sense judgment is that a hoagie is food and the conceptual analysis agrees. By contrast the claim, "An iron crowbar is food," is not rendered explicitly true because our (correct) common sense judgment is that a crowbar is not food. At best, the conceptual analysis can appeal to an explanation that a crowbar should, after all, be technically considered food because iron is nutritious, people do not think of it as food because it is hard to chew and digest in crowbar form.

I will now discuss two dimensions along which old-fashioned conceptual analysis can be relaxed. The first involves dropping the requirement that the analysis provide an explicit definition. Old-fashioned orthodox analyses have room to accommodate the vagueness of the target concept by way of the vagueness of the concepts in terms of which the explicit definition is formulated. But Frank Jackson (35) alone and with David Chalmers (7), for example, advocate permitting analyses that are vague in the degree of fit they make with the structure of necessary and sufficient conditions. This lowering of the bar is motivated by the recognition that our cognitive grip on some facts comes through a pattern recognition capacity rather than a rule checking capacity. For even the most mundane concepts, like our concept of the alphabetic character G, it is very difficult to write out an explicit definition of G, i.e. a specification of which glyphs count as clear cut instances of G holding across a wide variety of typefaces. Despite the seeming lack of an explicit rule for G-ness, people have widely shared opinions about the extension of G, with some vagueness at the borderline. This suggests that our concept of G exists by virtue of a shared capacity to recognize exemplars of G and to accommodate for variations from the exemplars. So, the more relaxed version of orthodox analysis Jackson defends allows the advocate of some analysis to incorporate fudge factors in the necessary and sufficient conditions that implicitly rely on our shared pattern-matching capacities. There is a danger that by permitting this kind of latitude, the conditions of adequacy will not be principled enough to differentiate between informative analyses and those that are mere data fitting, but this by itself is not a problem for conceptual analysis. We have good enough pattern-matching capacities that allow us to distinguish between trivial and informative analyses, even given the flexibility provided by this relaxation. So, I am in full agreement with Jackson on the advisability of this relaxation of the standards for an adequate conceptual analysis.

Old-fashioned orthodox analysis can be relaxed along a second dimension specifying how strictly the analyzed concept must fit the initial platitudes. The need to respect common sense judgments may seem like a clear enough criterion for an acceptable analysis, but proponents of orthodox analysis routinely take significant liberties in their conceptual analyses. In analyses of causation, for example, language pragmatics are often employed to explain away cases where the philosopher wants to proclaim some event as a cause even when regular folk do not. People often tend to think that *L*, the presence of primitive life forms on Earth millions of years ago, is not one of the causes of the French revolution. Philosophers usually count *L* as a cause because it has many of the signature characteristics distinctive of causation. It occurred before the French Revolution and was connected to the Revolution by way of a continuous stream of physical interactions. *L* was important in bringing about the Revolution in the sense that had there been no life on Earth millions of years ago, it is very unlikely French society would have evolved, much less had a revolution of the prescribed character. Ordinary causal discourse also distinguishes between foreground causes and other background conditions. People tend not to cite the presence of oxygen as one of the causes of the flame initiated by striking a match, but orthodox metaphysicians are generally happy to count it as a cause and explain away its lack of psychological salience as a feature of people's tendency to disregard standard background conditions or to focus their attributions of causation on changes to the status quo. The same can be said for many other factors playing a role in people's psychology of causation. In contemporary practice, the unassailable data that accounts of causation are expected to accommodate are philosophically regimented intuitions, not folk intuitions. Orthodox analysts are said to seek a "broad and non-discriminatory concept of causation" (41) or an "egalitarian" notion of cause (22), stripped of linguistic and explanatory pragmatics. Any analysis operating under standards of adequacy that have been relaxed along both dimensions, can be labeled a ***new-fangled orthodox analysis***.

Once new-fangled orthodox analysis has opened the door to various methods of explaining away mismatches between folk intuitions about a concept and the theoretically analyzed concept, it is unclear how people are expected to adjudicate between analyses that differ with regard to how many of the platitudes need to be rendered explicitly true rather than explained away as explicitly false but nonetheless reasonable. Consider the characterization provided by Collins, Hall, and Paul (9):

> It is clear enough—at least for present purposes—why someone interested in providing a conceptual analysis of our ordinary notion of causation should attend carefully to intuitions about cases. What we wish to emphasize is that even someone interested in "synthesizing" a new

> and potentially useful concept needs to heed these intuitions, else she risks cutting her project free of any firm mooring. More specifically, a reasonable and cautious approach for her to take is to treat intuitions about cases as providing a guide to where interesting causal concepts might be found. Thus, although the account can selectively diverge from these intuitions, provided there are principled reasons for doing so, it should not diverge from them wholesale. (p. 31)

In effect, new-fangled and old-fashioned orthodox analyses exist on a continuum where the new-fangled version is as lenient as possible regarding fit with folk intuitions while still being an orthodox analysis by insisting on reasonable (if imperfect) fit with the regimented intuitions. The authors proceed to elaborate on eight strategies for accommodating mismatches between theory and intuitions that allow the theory to count as successful. These involve the familiar maneuvers of explaining away discrepancies in terms of language pragmatics and accepting counterintuitive theoretical implications as a minor unfortunate side effect in order to gain other benefits from the theory.

The Collins, Hall, and Paul characterization of acceptable analyses does not make clear how it would differ from what I call 'empirical analysis'. Imagine a food scientist who has figured out everything important about nutrition and expresses these nutritional facts in terms of a regimentation of the folk food concept called 'nutrient'. Does her theory count as having only "selectively diverged" from the clear intuition that an iron crowbar is not food, or does that count as a principled deviation? Certainly no food scientist has ever explicitly explained away even a small fraction of the discrepancies between 'food' and 'nutrient'. Does that show that our complete knowledge of nutrition has nonetheless failed to tell us anything about food, or are the arguments that explain away the non-food status of iron crowbars so obvious that no one needs to provide them all explicitly? Is to "diverge wholesale" from the folk intuitions a matter of having too small a fraction of the folk platitudes come out explicitly true rather than explicitly false but pragmatically explainable? Or is it instead just the uncontroversial truism—accepted by empirical analysis—that in order for a conceptual analysis of *X* to be relevant to *X* rather than some other topic, it must be closely enough related to the folk platitudes regarding *X*?

Although the quoted characterization of conceptual analysis from Collins, Hall, and Paul is compatible with empirical analysis, I believe an examination of the practices of orthodox conceptual analysts supports the conjecture that they are engaged in an activity significantly different from empirical analysis. Except for some quibbles, everything I have said so far about empirical analysis is also compatible with Jackson's (35) clarification (pp. 30–36) of conceptual analysis. Nevertheless, an examination of Jackson's practical applications of his version of conceptual analysis, e.g. his discussion of color, makes clear that he is interested in locating a concept of color that fits folk intuitions about color much more closely than the chemist's concept of metal oxides fits folk intuitions about rust. The same holds for discussions of causation by Collins (8), Paul (53), and earlier work of Hall, e.g. (20)(21), though Hall's (22) paper, "Two Concepts of Causation," departs from the requirement that an adequate conceptual analysis needs to systematize all the platitudes regarding causation as a single regimented concept.

Empirical analysis, I think, differs from orthodox analysis in two key respects. First, empirical analysis is maximally liberal with regard to fit with the naive platitudes along both dimensions. Like Jackson's and Chalmers' version of orthodox analysis, a successful empirical analysis does not require explicit definitions of the analyzed concept. Empirical analysis is also at the most liberal extreme of what is allowed by the explicit recommendation given by Collins, Hall, and Paul because the folk intuitions are mere starting points for the exploration of a regimented concept that only needs to be close enough to the original platitudes concerning X so that it is not misleading to say that the empirical analysis is an analysis of X.

Second, empirical analysis includes an extra principle that guides movement away from the naive platitudes. Empirical analysis takes our naive platitudes concerning X as a starting point for isolating some empirical phenomena. Then, we seek a scientific explanation for those phenomena, honing the concepts used in the explanation as much as needed to optimize the overall quality of the explanation, including how it comports with other background theories we accept. Whatever concepts result from this optimization constitute the empirical analysis of X. An empirical analysis often results in some of the original platitudes being discarded as irrelevant to the analysis, and there is no demand that the final regimented concept make the platitudes come out explicitly true. While orthodox analysis is forever tethered to the initial platitudes, empirical analysis encourages us to abandon them whenever their literal truth would make the overall conceptual scheme suboptimal.

This explicit characterization of the difference between orthodox and empirical analysis can only communicate so much. An adequate grasp on the essential difference can only come by looking past vague statements of principle and examining how orthodox and empirical analysis work in practice. When we do this we will see that although there is no principled distinction between the kind of conceptual fit permitted by empirical analyses and that permitted by new-fangled orthodox analysis, there is enough of a practical difference to group the new-fangled and old-fashioned orthodox analyses together and identify their methodology as significantly different from empirical analysis. Let us now examine the practices of the orthodox analysts by looking at the metaphysics of causation.

13.1 The Orthodox Metaphysics of Causation

The orthodox investigation of causation more or less seeks a single structure simultaneously optimized for two tasks. The sought after egalitarian notion of cause is supposed to vindicate our ordinary causal talk by making central elements of this talk explicitly true, not just an understandable interpretation of reality, and it is supposed to be integrated with related metaphysical concepts like laws, counterfactuals, influence, control, dispositions, powers, time, etc. What defines an investigation of causation as orthodox is that the standards for judging its adequacy demand that an account relate causation to other interesting concepts in a principled manner (in the sense of not being just a data-fitting exercise), and that it adhere to the STRICT standards, and that a theory's pronouncements adhere closely to how people think about particular instances of causation as well as how they construe influence and causal dependence and express such commitments in ordinary language. In brief, the orthodoxy demands that accounts of causation be principled, STRICT and closely match core platitudes.

My account of causation also tries to explain causation in the world and our psychology of causation, but it does so with two distinct projects with different standards of adequacy. An adequate account of causation must be principled and held to STRICT standards, but need not accord closely with folk judgments about individual cases. Theorizing about our psychology of causation ought to accord closely with folk judgments about individual cases, but need only satisfy the RELAXED standard of theoretical adequacy. So, my project in this volume can be thought of as an attempt to produce two empirical analyses—one of causation and one of the psychology of causation—that stand as a replacement for the unified orthodox conceptual analyses that are routinely produced by metaphysicians of causation.

In orthodox conceptual analysis, causation in the world and our psychology of causation are unified in a single notion of cause that is to be investigated under standards that would be acceptable to both metaphysics and psychology. Because the metaphysician investigating causation is typically concerned with causation as something putatively out there in the world and construes it as fairly closely connected to fundamental reality, it makes sense that her theory of causation be held to STRICT standards. After all, it is standard practice in metaphysics generally to adopt STRICT standards, and STRICT standards are appropriate for any theory of structures playing a fundamental or nearly fundamental role. Because the philosophical investigation of causation typically takes causation to be the kind of relation implicit in ordinary causal claims, it makes sense that one's theory of causation should be required to match our suitably regimented pre-theoretical intuitions about causes. However, because an orthodox analysis tries in effect to systematize a lot of psychological data under a standard that is much more demanding than is the case in the rather high level psychology appropriate to judgments of causation, it makes such analyses very hard to complete successfully. It is hardly surprising that orthodox analyses have such a poor track record of systematizing all the psychological data under STRICT standards.

There are two subsets of the causation literature that illustrate the style of inquiry that sets orthodox metaphysics apart from a more scientifically oriented metaphysics. Both illustrate the peculiar activity of mixing psychological and linguistic concerns with metaphysical concerns.

13.1.1 CAUSATION AND ORDINARY LANGUAGE COUNTERFACTUALS

There is arguably an important connection between causation and counterfactuals. A ***counterfactual*** is a counterfactual conditional, a claim about what would be true if certain (typically non-actual) circumstances had obtained. Linguists, logicians, psychologists, and philosophers of language investigate the logic, syntax, and semantics of natural language conditionals, including counterfactuals. People often have "clear intuitions" about counterfactual claims regarding particular causal happenings as well as about general inference patterns involving counterfactuals. One of the marks of the orthodox approach to the metaphysics of causation is that it takes seriously the idea that the logic of ordinary language counterfactuals and intuitions about particular counterfactuals provide an important data set, the explicit truth of which a theory of causation needs to be compatible with. To the extent the concept of causation is part of a larger conceptual scheme involving influence and counterfactual dependence, pretheoretical intuitions about which counterfactuals sound naturally correct become part of the overall set of platitudes one's conceptual analysis of causation needs to match.

From the perspective of empirical analysis, there is a rather straightforward skeletal account of the proper relation between causation and ordinary language counterfactual claims. There is some objective structure in reality that ultimately accounts for the existence of effective strategies and important regularities about effective strategies. This structure ultimately grounds some of our causal talk and some of our counterfactual talk. Although there does need to be some adequate account of effective strategies, our ordinary causal talk and ordinary counterfactual talk might be explainable in a way that does not require the platitudes to be explicitly true or organizable into a STRICT system.

Consider the example known as Morgenbesser's coin (63) p. 26 fn. 33. Suppose the world is governed by fundamentally chancy laws and that there is enough randomness in the microscopic world so that the fifty percent chance that an ordinary coin flip lands heads is overwhelmingly due to fundamental chanciness, not to our ignorance of the microscopic details of the setup. A coin is flipped and when it is airborne, Jane bets heads, and the coin lands tails. When people are told of such stories they tend to agree with the counterfactual, "If Jane had bet tails, she would have won." Orthodox theories, e.g., (51) (59), tend to take such folk opinions as truths that need to be entailed by any adequate analysis, not merely as practices that are understandable as folksy approximations of some deeper structure that is relatively far removed from the explicit content of the counterfactual claim. Again, not all folk intuitions are sacrosanct according to new-fangled orthodox analysis, but to the extent causation is interpreted as part of a larger theory that includes counterfactual claims, the orthodoxy tries to impose some burden of explanation on theories that deny the Morgenbesser counterfactual.

13.1.2 CULPABLE CAUSES

The second example of how the orthodox metaphysics of causation is a mixture of psychology and metaphysics applies to virtually every theory of causation. Although orthodox metaphysical accounts of causation can have different overall goals, one of the tasks of any orthodox account is to identify non-trivial rules for which events count as causes, given not-too-causally-loaded information about the laws of nature and the history of occurrent facts. When we cite instances of causation—a whale breach causing a splash or an accident with a cactus causing pain—we intend to draw special attention to a small portion of the universe as being important to the effect. These events are what in ordinary language and in philosophical discourse are called "the causes" of the effect. Orthodox theorizing about causation takes as its task explaining rules for what makes something count as one of the causes. These are called ***singular*** causes because they are the events that (allegedly) cause the particular effect in that one fragment of the world's history. We can contrast singular causation with ***general*** causation, which concerns what happens generally across many fragments of history, e.g. that whale breaches cause splashes and accidents with cacti cause pain.

The sought-after singular causes are typically not fantastically detailed physical states but are intended to be the kind of events people tend to cite when asked about the causes of some particular event, e.g. the launching of the ship, the loss of a tooth, the increase of gross domestic product in the fourth quarter of 1968. From here on, I will refer to such events as ***mundane events***. Orthodox accounts of singular causation focus on relations among mundane events, though they typically allow that causal relations can exist among events that are physically sophisticated, e.g. the total microphysical state existing on an infinitely extended time slice. Because these sophisticated kinds of events might play a role in singular causation, it is valuable to distinguish the kind of singular causes that are mundane events. A ***culpable cause*** of some event E is an event that counts as "one of the causes of E" in the sense employed by metaphysicians who study causation. 'Culpable cause' is not a technical term but merely a label for the "egalitarian" (22) or "folk attributive" (27)(32) notion of cause that orthodox metaphysicians seek when they ask, "What are the causes of (the singular event) E?" I emphasize that 'culpable cause' is my proprietary expression[19] introduced to reduce confusion about what 'cause' by itself connotes. When I claim that people have intuitions about causal culpability, I do not mean that ordinary people understand the expression 'causal culpability', but merely that people have implicit beliefs about singular causation among mundane events. It is that implicit concept that I am labeling as 'culpable cause'. Two further qualifications can be made at this point. First, culpable causes are so named because they are events that are blameworthy for the effect, but the terminology is not meant to imply that our intuitions about the relevant notion of singular cause absolutely perfectly matches our intuitions about how to attribute causal blame. Second, there is an ambiguity in the expression 'a cause of E'. It could mean 'one of the causes of E' or it could mean 'something that caused E'. These are

[19] The term 'culpable cause' has been used previously by Mark Alicke (4) to designate something altogether different: the psychological effect of perceived moral blameworthiness on judgments of causal impact.

not equivalent. When Guy won the lottery, his purchase of the lottery ticket was one of the causes of his winning but it was not an event that caused him to win. 'Culpable cause' refers to the 'one of the causes' disambiguation.

To summarize, according to the orthodox metaphysics of causation,[20] any adequate account of causation must provide an acceptable account of culpable causation. A successful account must provide principled, STRICT rules for when a given event is culpable for some chosen effect, and these rules must accord with an acceptably large number or fraction of platitudes concerning culpable causes.

Orthodox conceptual analysis is legitimately described as orthodox because virtually all the academic literature on causation to some extent or other assumes the STRICT standards of adequacy and the attention to intuitions about culpable causes that the orthodoxy demands. The orthodoxy certainly includes the classics: Mackie's (46) inus account, Lewis's (41) and (44) counterfactual dependence accounts, and Suppes' (65) probabilistic dependence account.

There is a fraction of the contemporary philosophical literature that at least superficially disavows orthodox conceptual analysis, especially Hausman's (28) account in the probabilistic dependence tradition and Dowe's (10) account in the transference tradition. They each provide useful discussions of how their projects differ from old-fashioned orthodox conceptual analysis, but despite their explicit rejection of orthodox conceptual analysis, in practice they exert significant effort to account for intuitions about culpable causes. Hausman is explicit about seeking to make "paradigm causal claims" turn out true (p. 10) and Dowe develops, in his Ch. 7, an account of a causation-like concept meant to vindicate some intuitions about culpable causes. It is unclear whether their accounts are intended to be held to (what I have identified as) STRICT standards, but nothing suggests that either author thinks of the rules governing the correct identification of culpable causation as being psychological heuristics that are metaphysically dispensable.[21]

One recent trend in the study of causation that is not closely tied to the orthodoxy is the causal modeling tradition based on the work of Spirtes, Glymour, and Scheines (64) and Judea Pearl (54). Although their interest in causation is not squarely metaphysical, the models have been appropriated for metaphysical purposes, e.g. (30) (71) (24) (25) (49) , in order to develop in some cases, STRICT theories of culpable causation. Although I believe attempts to extract a STRICT account of culpable causes is unnecessary, the use of the causal modeling approach to understand causal generalities is not a target of any criticism in this book. Although my account of causation is not based on the structures invoked by the causal modeling approach, I do accept that causal modeling is a useful scientific practice and that my account of causation would

[20] I am defining 'orthodox metaphysics of causation' so that this claim is true by stipulation, but I do believe that the actual practices of philosophers who publish on the subject of causation demonstrate that most of them believe an adequate account of causation requires a STRICT, principled account of culpable causation.

[21] Dowe is clear in his Ch. 6 that intuitions about omissions can be satisfactorily vindicated without his account needing to make it explicitly true that the causation by omission is real causation.

be inadequate if it could not make sense of its utility. While I do not have the space to present an adequate comparison of my account with causal modeling approaches like that of Woodward (71) and Sloman (62), I can make two brief comments. First, my later discussion of counterfactual conditionals and backtracking is intended to vindicate talk of the 'intervention counterfactuals' that are invoked by causal modeling approaches. Thus, I see my account as complementing causal modeling approaches to causation. Second, my primary reason for preferring my account of the metaphysics of causation to any account based on the causal modeling approach is that I believe my account can help to elucidate why causation appears to be future-directed and why there is no genuine causal backtracking, whereas causal modeling approaches typically build these features of causation into their models, leaving them unexplained.[22]

13.2 The Orthodox Metaphysics of Culpable Causation

When theorizing about culpable causes, philosophers like to play the following game. Someone offers a theory of causation providing rules for when it is correct to say C was a cause of E and when it is correct to say C was not a cause of E. The theory is allowed to remain silent on some cases and can relativize its pronouncements to parameters the theory specifies. Then, opponents attempt to formulate counterexamples by identifying some scenario where a high enough level of agreement can be secured among causation experts about whether C was a cause of E based on opinions that are not too theoretically informed.

I will now discuss the three conditions of adequacy that the orthodox metaphysics of causation places on any account of culpable causation. First, any successful account must closely accord with philosophically regimented (but not too theoretically informed) intuitions about culpable causes in (preferably realistic) test cases. Second, any successful account must provide a principled unification of what is common in all (or nearly all) cases of culpable causation. Third, a successful account must be STRICT, i.e. free of conflicts.

13.2.1 CLOSE FIT TO PSYCHOLOGICAL DATA

Consider an unremarkable situation in which a match is intentionally struck, which generates a flame, which is then used to ignite a fuse that burns until it launches a rocket at time t. Also, after the fuse was lit, a bystander made the decision to launch the rocket herself at t by walking up and directly launching the rocket with an electric starter, but after seeing that the fuse was going to launch the rocket anyway, she changed her mind and just stood there watching the launch. Here are some exemplars of the kinds of intuitions about culpable causes that are generally considered uncontroversial truths that any adequate analysis must agree with.

[22] See Weslake (68) for an explanation of why this is so.

- (Irreflexivity) The rocket launch was not a cause of the rocket launch.
- (Asymmetry) The rocket launch was not a cause of the match being struck.
- (Preemption) The bystander's decision to launch the rocket was not a cause of the rocket launch.

There is nothing remarkable about the particular details of this one example. All three propositions represent principles that hold generally across a wide variety of commonplace instances of causation.

I take these three propositions as uncontroversial data that must turn out to be explicitly true on any adequate orthodox account of causation. According to Collins, Hall, and Paul (9), orthodox analysis does permit intuitions about cases like these to be ignored if there is some "overriding reason." The problem with allowing such maneuvers as part of the orthodox metaphysics of causation is that the only difference separating a liberal version of the new-fangled orthodox analysis and empirical analysis is that empirical analyses are not required to make such cases turn out to be explicitly true but only understandable in light of heuristics that obey RELAXED standards. Because these three examples happen to be extremely uncontroversial among orthodox metaphysicians, if a metaphysician is willing to deny their explicit truth, I would begin to lose my confidence that he is genuinely operating under the orthodox standards of theoretical adequacy. In principle, someone could be orthodox yet deny the truth of these claims so long as he holds steadfast to the explicit truth of enough other claims concerning culpable causes, but I suspect these three intuitions are so central to the core idea of causation that if there are good enough reasons to explain them away, there are probably good enough reasons to explain away truths about more controversial causal principles such as transitivity. So, in order to maintain some principled distinction between orthodox metaphysics and scientific metaphysics, while being as generous as possible to the new-fangled analyst, I will just stipulate that what I am calling the project of orthodox metaphysics of causation includes the task of providing an orthodox conceptual analysis of culpable causation such that all three principles come out explicitly true, and I will leave open whether an orthodox metaphysician needs to render any other intuitions explicitly true. If the selection of these three principles and no others sounds too arbitrary, I agree. But the rhetorical predicament I face is that the orthodox metaphysics of causation employs the inherently shifty methodology of new-fangled orthodox analysis. Without drawing some line to distinguish new-fangled orthodox analysis and empirical analysis, it is difficult to communicate their essential difference. After getting the gist of my overall argument against this somewhat arbitrarily chosen target, it should be clear to readers how to adjust the argument if some orthodox theorist chooses to deny one or more of these three principles.

In the orthodox metaphysics of causation, one treats clear intuitions like these three principles as unassailable facts about the nature of causation, whereas some other platitudes concerning causation may be brushed off as merely the result of language pragmatics, not genuine truths about the egalitarian concept of causation that the orthodox theorist accords metaphysical prominence. This distinction turns out to be unstable, however, because the pragmatics that the orthodox metaphysician himself uses to explain away some intuitive truths are easily turned against the three signature principles. Let us now review how easily their privileged status can be called into question by providing adequate explanations for them in terms of pragmatics.

13.2.1.1 Irreflexivity

Lewis (44) advocates irreflexivity, stating that the cause "C and [effect] E must be distinct events—and distinct not only in the sense of nonidentity but also in the sense of nonoverlap and non-implication. It won't do to say that my speaking this sentence causes my speaking this sentence; or that my speaking the whole of it causes my speaking the first half of it, or vice versa; or that my speaking of it causes my speaking it loudly, or vice versa." (9), p. 78[23]

Contrast the rejection of self-causation and causation-of-parts with standard practices in the sciences. An engineer who is interested in understanding the rotation of material objects is well served by group theory, the branch of mathematics most useful for characterizing physical symmetries. The group SO(2) is a mathematical structure for modeling the relations between all the possible rotations an object can undergo in a two-dimensional plane. The elements of this group can be represented by real numbers. The number θ corresponds to a counter-clockwise rotation by θ radians. Negative numbers correspond to clockwise rotations and the zero rotation corresponds to no rotation at all. Applying the principle that an analysis of the concept of rotation must hold to clear opinions in the sense Lewis advocates requires that we reject any analysis of rotation that counts a zero degree rotation as a rotation. What could be a clearer instance of a non-rotation? The reason zero rotations are included in the group-theoretic concept is that it greatly simplifies the theorems concerning relations among different kinds of rotation. For example, we would like to be able to say that the composition of any two rotations is itself always a rotation, but we cannot state that claim with optimal simplicity if zero rotations are forbidden because a rotation by θ and then by $-\theta$ amounts to a net non-rotation. Mathematicians understand the zero rotation as a trivial rotation rather than something that is not a rotation at all. This consequence of orthodox analysis—that the SO(2) account of rotation is refuted by the clear intuition that to rotate by zero degrees is not to rotate at all—highlights the key problem with the orthodox analyst's devotion to clear intuitions. It sacrifices conceptual optimization merely for the sake of making it explicitly true that rotations by zero degrees are not rotations.

What goes for rotation goes just as well for causation. One can easily treat self-causation as a case of genuine causation, albeit a trivial one. On just about any standard theory of causation, the event E has the right kind of relationship to itself to count as causal. E is a condition that lawfully and non-superfluously necessitates E. E counterfactually depends on E's occurrence. E is physically connected to E via a (trivial) physical process. E raises the probability of E from what it would have been without E. It is also easy to see why it is reasonable for us to think of an event's relationship to itself as always non-causal: such relations always exist regardless of the

[23] Notice that Lewis cleverly frames the issue in terms of what would be wrong to say, which permits the interpretation that it might be merely pragmatically misleading rather than explicitly false that events cause themselves.

event and regardless of the laws of nature.

Not only is it acceptable to model self-causation as trivial causation rather than as a lack of causation, there is a good reason for doing so. Existing orthodox accounts already need explanatory pragmatics to account for why we do not cite causes that occur a trivially short amount of time before the effect, and these pragmatics automatically cover the case where the trivially short amount of time is zero. For illustration, suppose Jill is sleepy and goes to bed early in the evening and stays in bed until the late morning without anything remarkable happening. Let E be Jill's being asleep in bed at exactly midnight. What are the causes of E? One of the causes is C, the fact that she is asleep in bed exactly 10^{-50} seconds before midnight. C is not the kind of cause one would normally cite when providing a causal explanation of E because on the time scales relevant to a causal explanation of human behavior it amounts to little more than a restatement of the event to be explained, but it does count as a cause according to the rules provided by prominent accounts so long as they permit reference to the brief events I described. Whatever story one employs to explain away the disutility of citing C will likely extend to self-causation because in the limit as time t approaches midnight, the event of Jill being in bed at t becomes ever more useless for explaining E. If we take the orthodox approach and require that causation is irreflexive, then we in effect explain the wrongness of citing her condition at midnight as a cause of E in two different ways depending on the fine difference between the state at precisely midnight and states arbitrarily close to midnight. At precisely midnight, her condition is not to be cited as a cause of E because it is false that E causes E, but at any moment just before midnight, her condition is not to be cited as a cause of E purely on the pragmatic grounds that it is informationally unhelpful given typical human concerns. If we ignore the orthodox approach by adopting the unconventional hypothesis that events do cause themselves, we can say that pragmatics governs the wrongness of citing her condition throughout, so that there is not a discontinuity in the nature of the explanation that depends on the fine distinction between midnight and just prior to midnight.[24]

13.2.1.2 Asymmetry

If it is true—as I demonstrate in the chapter on causal asymmetry—that influence directed toward the past never has any practical utility, then that automatically provides a pragmatic explanation for why it is reasonable not cite to events after E as causes of E even if they have other signature features of causation. An advocate of any of the traditional approaches to causation—e.g. inus accounts or counterfactual accounts or probability-raising accounts—can argue that even if some of the events that occur after E do technically cause E, our instinctive judgment that they are not causes can be explained away in terms of our having incorporated into our instinctive concept of a (culpable) cause, the general uselessness of past-directed influence. That is, because it is always useless to try to cause events in the past, we think that there are no causes of past events.

[24] Astute readers might want to counter that the asymmetry of causation requires that Jill's condition at an arbitrarily small time after midnight count as not a cause at all, so that we are left with a discontinuity regardless, but this can also be explained away, as discussed in the next subsection.

13.2.1.3 Preemption

Another instructive illustration of orthodox analysis is its treatment of preemption. The decision of the bystander, D, to launch the rocket was a cause of the rocket launch, E, in the sense of being one of the events that played a part in the overall physical development of the world toward the launch. It also raised the probability of the launch. Suppose we accept, contrary to received wisdom, that D is a genuine cause of E. We can explain why people have the intuition that D is not a cause as follows: In the vast majority of cases, when there is causation from an event C to a later E, there is a continuous physical evolution of the world from time t_C to t_E such that whatever difference C eventually makes to E is delivered by way of some physical differences in the intervening times. In our world, so far as we can tell, there are no nomic connections that leap over spans of time. Furthermore, because we often glean useful information about how the world works by observing patterns and tracing back from E through whatever physical patterns we construe as causal processes, we place a lot of practical importance on those causes that can be found by tracing back in time from E. If all that is correct, we have a simple explanation for why we do not identify the bystander's decision as a cause of the launch: The usual pattern of features we would expect if the bystander were causing the rocket to launch did not occur. We would expect things during that time span to exemplify a physical transition throughout the stages of a kind that is recognizably causal in the sense of matching what we think of as prototypical cases of a decision like D leading to a rocket launch. Nevertheless, nothing about the lack of the right kind of process prevents us from claiming correctly that D was a genuine cause of the launch although it is not recognized as such by folk judgment because it did not leave the usual indications that we use to identify causes. Of course, one could complain that identifying D as a genuine cause does not capture the relevant notion of cause that the orthodox analyst is seeking, but this is exactly the tenet being questioned. Why is *that* notion of cause the one that needs to be enshrined in the metaphysics as genuine causation rather than some more liberal notion whose lack of psychological salience is explained away? The orthodoxy's only answer is that folk do not cite D as a cause when presented with such scenarios and that folk intuitions are the touchstones of adequate analysis. Because the orthodoxy permits some recalcitrant folk intuitions to be explained away, that calls into question how principled the egalitarian notion of cause really is.

The point of these examples is not to settle whether any one particular platitude about causation is best construed as 'explicitly false but pragmatically understandable' rather than as 'explicitly true'. Their purpose was to emphasize that plausible pragmatic explanations are available for even the most uncontroversial clear intuitions that orthodox analysts hold dear. This in turn raises the question of whether there is something special about causation that requires analyses of causation to make the central folk intuitions explicitly true, even though for understanding rotation or food, it is perfectly acceptable for the theoretically reformed concept to account for folk usage in ways other than explicit truth.

13.2.2 PRINCIPLED ANALYSIS OF CULPABLE CAUSES

One feature that makes the orthodox analysis of causation a project in metaphysics rather than armchair psychology is that a proper analysis is required to provide a principled account of what is common to all cases of causation. Suppose a psychologist offers a theory of causation consisting of a list of 8 exemplars of the cause-effect relation and 13 exemplars of the lack of a cause-effect relation. The theory says C is a cause of E if and only if the situation where C and E happens is closer to one of the positive exemplars than to any of the negative exemplars, closeness being judged by one's own intuitive off-the-cuff assessment of similarity. A theory of this form might make for an interesting psychological theory and might even accrue empirical support if our causal reasoning is based less on rules than a pattern-matching capacity. But from the perspective of metaphysics, it fails to capture what is similar in all the cases of causation in an interesting way. Such theories come across as merely fitting the data, whereas the metaphysician is interested in a theory based on principles, something more closely resembling necessary and sufficient conditions.

13.2.3 STRICT STANDARDS FOR ACCOUNTS OF CULPABLE CAUSES

Another feature that distinguishes orthodox analyses of causation from psychology is the expectation that they are to be held to STRICT standards of consistency, as defined in the introductory chapter. An easy way to see the difference between the psychologist's standards for theoretical adequacy and those of the metaphysician is to consider the following toy theory of causation.

1. An event C is a cause of E if and only if C raises the probability of E.
2. An event C is a cause of E if and only if there exists a chain of probability-raising relations going from C to E.

This conjunction of rules might be faulty for multiple reasons, but let us focus just on realistic possibilities where the rules conflict. In an example (65) attributed to Deborah Rosen, a golfer's slice, C, lowers the probability of a good shot, E, and so is not a cause of E according to the first rule, but the slice does raise the probability of hitting a tree, which in turn raises the probability that the ball will bounce back in a better position making C a cause of E according to the second rule.

By the RELAXED standards of psychology, it is acceptable for a theory to claim that people employ both rules as heuristics for assessing causation despite their genuine conflict. The psychological theory could make a further prediction that in cases of conflict, people will become less sure of their judgments or perhaps that some other secondary factors come into play to nudge a person into favoring one rule over the other. There might also be priming effects or context effects or interactions with people's attention mechanisms, etc. Although it would be nice for a psychological theory to pin down all such factors, it is plausible that as one improves a theory to make it increasingly precise about which rule we implicitly select, that will require an increasing quantity or specificity of parameters, so that the theory's predictions and explanations become increasingly complicated and thus decreasingly valuable. By the ordinary RELAXED standards of psychology, having multiple conflicting rules for what events count as causes can be acceptable even if there is no further account in the theory of how to resolve (for all realistic circumstances) which heuristic is operative.

But from the point of view of metaphysics, conflicting rules are unsatisfactory as an account of causation. In metaphysics, one is thinking of the causes as some element of external reality. A theory that provides conflicting pronouncements about whether *C* is a cause and provides no further device to settle which rule is applicable and fails to relativize the incompatible facts to parameters that would remove the conflict, is in effect stating that its model of world is inconsistent, which is uncontroversially unacceptable. One of the crucial standards by which orthodox metaphysical theories are to be judged is that their rules for causation need to be consistent. Furthermore, one is not allowed to save the inconsistent rules merely by adding a hand-waving qualifier that says, "In some cases the first rule holds and in others the second rules holds." One is obligated, according to the implicit standards of orthodox metaphysics, to provide parameters such that there is at most one answer to whether *C* is a cause given the parameter settings.

My empirical analysis of causation is meant to be held accountable to the STRICT standard. Where my account differs from the orthodoxy is that I hold that rules about culpable causation are not metaphysical rules but psychological rules. Thus, for me, the above pair of rules should not be tossed aside because they conflict, for I interpret them merely as psychological heuristics governing our folk cause concept, not rules governing the structure of causation itself. The kinds of causal concepts that do play a role in the metaphysics of causation such as determination, probability-fixing, probability-raising, and influence, do need to be held to STRICT standards, but not the notion of culpable cause.

For purposes of discussion, I hereby stipulate that the orthodox metaphysics of causation is identifiable with the following standards for what kind of theory counts as successful. On the one hand, orthodox metaphysical theories of causation are expected to provide principled rules for something's being a (culpable) cause and these rules must obey the STRICT standards for consistency. On the other hand, its concept of causation is expected to closely match psychological data concerning judgments of causal culpability. This presumably also includes rendering the three principles—irreflexivity, asymmetry, and preemption—as explicitly true.[25]

[25] Remember that in principle, someone providing a new-fangled orthodox account of the causation concept could eschew these three principles in favor of defending some alternative clear intuitions, but in order for the standards of new-fangled orthodox analysis to avoid being so weak as to effectively collapse into those of empirical analysis, there needs to be at least some minimum basis of clear intuitions that are readily recognized as such. I selected these three as defining the minimum basis of the orthodox metaphysics of causation because they are extremely uncontroversial claims, when understood as applied to ordinary cases like that of the rocket launch. I am not assuming that the example statements I called 'irreflexivity' and 'asymmetry' are fully general principles.

13.3 The Orthodox Metaphysics of Causation is Unneeded

The ideal food scientist, who has figured out everything there is to know about human nutrition and recognizes that the primary reason we have a food concept is that it gives us a cognitively efficient grasp of nutrients, will be unfazed by the philosopher's "counterexample" that earthworms are nutritious but not food according to common sense. Nor will she be flustered by the philosopher's complaint that despite all her work, she has not really been studying food because—as many counterexamples demonstrate—being a food is obviously not equivalent to being a nutrient. And rightly so. Such attacks on the nutrient concept are entirely irrelevant to the quality of the explanations provided by food scientists and to the applicability of food science to our understanding of food.

Analogously, discrepancies between scientifically honed causal notions and folk intuitions concerning culpable causes and Morgenbesser's coin are not automatically counterexamples to metaphysical claims regarding causation and related notions of influence and counterfactual dependence.

Putting the conclusion in more general terms, empirical analysis is defensible because it is the form of conceptual analysis routinely employed in well-functioning sciences and has earned its keep because numerous sciences have implicitly employed empirical analysis to a successful end. The empirical analysis of causation in particular is defensible because causation is presumably a subject matter amenable to science, just like its various special cases: gravitation, combustion, erosion, etc.

What the orthodox metaphysics of causation attempts to accomplish is to find an optimal notion of causation that on the one hand is principled and STRICT and on the other hand closely fits the psychological data. What my account does is to replace this project with two empirical analyses. The empirical analysis of causation is principled and STRICT but does not closely fit the psychological data. The empirical analysis of the psychology of causation is principled and closely fits the psychological data but only satisfies the RELAXED standard. This pair of analyses accomplishes what the orthodox approach attempts to do in a single analysis, but because it segregates the needed concepts into a metaphysics part and a psychology part, it is able to optimize the metaphysical concept in accord with the demands of metaphysics and the psychological concept in accord with the demands of folk intuition. It is thus able to achieve greater optimization without losing anything important.

For my particular account, the rule determining what belongs in the metaphysics of causation and what belongs in the psychology of causation is this: Whatever is relevant to the explanation of effective strategies is part of the metaphysics of causation. Whatever is irrelevant to the explanation of effective strategies but bears on our folk notion of causation is part of the psychology of causation. Because my explanation of effective strategies employs relations of nomological determination and probability-fixing and probability-raising, it is incumbent on me to ensure that there are no conflicts in my rules about them. The reason my account of culpable causes only requires the RELAXED standard is that they turn out to be irrelevant to the explanation of effective strategies, as discussed in the chapter on culpable causation.

13.4 Criticism of the Orthodox Metaphysics of Causation

For the purpose of defending my own theory, it suffices that an informative metaphysical theory of causation can be constructed that eschews the strictures imposed by the orthodox metaphysics of causation. But one can go further, I think, and reject the idea that an orthodox metaphysics of causation tells us something interesting about causation that we cannot get from a pair of empirical analyses.

13.4.1 UNCLEAR MOTIVATION

One question that has never been satisfactorily answered is, "What is the purpose of an orthodox analysis of causation?" Explanations of why we need some or other conceptual analysis are commonplace, e.g. (55), p. 65: we need to know what we are talking about. But the relevant question is, "Why do we need an orthodox conceptual analysis of causation rather than a pair of empirical analyses, one directed at causation itself and the other at our psychology of causation?" A desire for a theory of our ordinary notion of causation is reasonable enough, but that is what a psychological theory can provide. Why does the orthodox metaphysician of causation insist that such a theory must satisfy the STRICT standard when the relevant kind of psychological theories are reasonably held only to RELAXED standards? A desire to understand why the world behaves in its paradigmatic causal way is understandable as well, but why must the concepts optimized for understanding that aspect of nature closely hew to folk opinions about culpable causes?

The few explicit defenses one can find of the orthodoxy do not sufficiently address the question:

> [The goal of new-fangled orthodox analysis is to provide] a cleaned up, sanitized version of some causal concept that, though it may not track our ordinary notion of causation precisely, nevertheless can plausibly be argued to serve some theoretical purpose....
> Obviously, someone who pursues this...aim ought to say at some point what such purposes might be. But we think that she is under no obligation to make this clear at the outset. On the contrary, it strikes us as a perfectly appropriate strategy for a philosopher working on causation to try to come up with a clean, elegant, theoretically attractive account of causation (or some causal concept), in the reasonable expectation that such an account will serve some, possibly as-yet undisclosed, philosophical or perhaps even scientific purpose....(9), 30–31

If the orthodox project were merely advocacy for the free play of ideas in the hope of eventually finding some useful notion, it would at worst be an inefficient method to produce a tangible good. In reality, though, orthodox analysts routinely attack other people's accounts of causation for inadequately addressing counterexamples drawn from the well of common sense. This raises the obvious question, "On what basis can a theory be rejected for inadequacy unless some constraints on the purpose of the account have already been adopted?"

While I have no decisive argument that the orthodox methodology cannot result in an adequate account of causation, there are good reasons to question the wisdom of following the orthodox approach to the metaphysics of causation. There is the overlong history of futility in playing the philosopher's game of trawling for counterexamples, both in the causation literature and in philosophy more broadly. But more specifically, there is a simple explanation for why an adequate orthodox account has been so hard to find. The twin goals of matching folk opinions closely and obeying the STRICT standard pull the analysis in opposite directions. It is much easier to secure a precise, principled account of causal concepts like determination, probability-fixing, and probability-raising if one does not need to worry about intuitions about culpable causes. And it is much easier to secure a principled account of folk intuitions about culpable causes if one is free to adopt a flexible interpretation of the various heuristics we use to identify which events are causally culpable.

There is undoubtedly some benefit to having a unified theory of causation simultaneously honed to serve some metaphysical role as well as to account for why we have our shared body of intuitions about culpable causes. Such a theory would provide some valuable conceptual economy. But the relevant question is whether the gain in conceptual economy is worth the loss in conceptual optimization. A screwdriver made out of a carrot would have clear benefits; it would be lighter than an ordinary screwdriver and you could eat it if hungry. But given the obvious tradeoffs, it is hard to believe an engineer could design a carrot-screwdriver that would not be significantly outperformed by just having a metal screwdriver and an ordinary carrot separately. Conceptual economy is worth something, but not much. If a theory were to invoke a sizable number of different versions of our causal concepts without a clear enough account of how they are related, that would be a good reason for complaint. Having two or three kinds of causation and a story about how they fit together hardly strains our cognition. But the cost of replacing two concepts optimized toward different ends with a single causal concept that is optimized toward both simultaneously is significant. Barring a stroke of fortune, the complexity of our psychology of causation demands tradeoffs between the degree of fit with common sense intuitions and the simplicity of the rules governing the application of causal concepts. The new-fangled orthodox analyst already admits this, in that the whole point of the egalitarian notion of cause is to idealize away opinions that result from explanatory pragmatics for the sake of a simpler account of causation. What's more, the difficulty in achieving a good fit with folk judgments while being simple and comprehensive is easily measured by the vast volume of material written on the subject of culpable causation. Once the practical necessity of these tradeoffs is accepted, there is room for different accounts to trade off the fit in different ways for different purposes. What is peculiar about the orthodox analysts' take on causation is their frequent insistence that there is one right way to optimize the concept.[26] I suspect that what explains this curiosity is that orthodox practitioners conceive of conceptual analysis more as an activity of conceptual exploration and discovery rather than conceptual engineering and construction.

13.4.2 CAUSATION IS A SCIENTIFIC CONCEPT

As I previously noted, one of the most uncontroversial things that can be said about causation is that rusting, radiation, photosynthesis, digestion, gravitation, combustion, erosion, and oxidation are all special cases of causation. Causation, furthermore, is just our generalization of all these special cases and others like them.

Consider the example of rust. There are well enough understood scientific methods for grouping together all substances with a similar chemical character to the substances we readily recognize as rust. I do not think it would be a credible challenge to the applicability of chemistry to rust to point out that some scientific precisification of rust, say 'metal oxide', is not coextensive with our folk conception of rust. One of the interesting things we have learned about rust is that it bears an important similarity to combustion. Naively, there is nothing in burning wood that seems similar to rusting iron, but in explaining how both kinds of processes take place and in systematizing the relevant concepts we find that it is useful to generalize rusting and combustion under the general category of oxidation. Furthermore, in the move to generalize and categorize the various kinds of oxidation, one does not suddenly shift methodology. Oxidation is studied using the same scientific methodology and empirical analysis used for investigating combustion and rust individually.

The tendentious upshot of the orthodox metaphysics of causation is that it in effect instructs us, "Do not study causation using the same methodology and empirical analysis that you use to study rusting, radiation, photosynthesis, digestion, gravitation, combustion, erosion, and oxidation. When you get to the level of generalizing what all these species of causation have in common, it becomes crucially important that your theory also adhere closely to what people on the street think about instances of causation. Sure, some allowances can be made here and there for your theory of causation to diverge from folk intuition, but you need to avoid too many divergences and you are obligated to explain away the discrepancies with principled arguments, lest you 'lose your moorings.' "

The challenge for the orthodoxy is to explain what makes causation special in a way that requires that its STRICT conceptual analysis must be moored closely to folk opinions about causation while the conceptual analyses of all the various species of causation need only match folk opinions in the loose way that is uncontroversially acceptable in science. It does no good to cite the greater metaphysical significance of causation, for empirical analyses of causation are required to deliver a principled, STRICT analysis of causation as well. All an empirical analysis lacks is that the aspects of our folk conception of causation irrelevant to the explanation of effective strate-

[26] Hitchcock (31) discusses numerous examples of such pseudo-debates.

gies are delegated to the psychology of causation where they are given a RELAXED treatment. If the empirical analysis of causation and the empirical analysis of the psychology of causation succeed together at providing a complete scientific explanation of effective strategies and a complete scientific explanation of why we have the naive causal concepts we have, on what grounds will the orthodox defender argue that these explanations do not tell us everything we need to know about causation?

13.5 Changing the Topic

Because virtually all extant analyses of causation are of the orthodox variety, one might wonder whether my empirical analysis is really a competitor to these analyses rather than just pursuit of an independent line of inquiry that is compatible with orthodox approaches. I think that my account ought to be seen as a competitor because, just like orthodox theories, it attempts to explain what is common among cases of causation, to identify central features exhibited in paradigmatic instances of causation that explain their commonalities systematically. I think the situation is analogous to the following hypothetical dispute. Suppose a late nineteenth century physicist has a project to identify those particles or fields that instantiate the gravitational force and interprets the term 'gravitation theory' such that it is a priori the study of the gravitational force. After Einstein produces his general theory of relativity, GR, the physicist could argue that Einstein's GR is not really a theory of gravity because GR asserts that there is no gravitational force. I do not think we would take this physicist's objections seriously because, regardless of the stipulation about what counts as genuine gravity, GR provides a superior account of the motion of bodies. In the same sense, although my investigation proceeds using a significantly different methodology, it is without any serious question a metaphysical account of causation.

13.6 Summary

Empirical analysis is the form of conceptual analysis routinely employed in the sciences. If attempting an empirical analysis of causation is wrong-headed, that is either because (1) empirical analysis in general is wrong-headed in which case we have much bigger problems than anything related to my empirical analysis of causation, or (2) something specific to causation makes it unsuitable for scientific inquiry, a claim which no one has adequately defended and which flies in the face of many successful empirical analyses of particular species of causation.

Orthodox analysis, including the new-fangled variety, is defective because it rejects conceptual analyses that should be considered entirely adequate. As many examples show, e.g. rotation, an empirical analysis can be unimpeachable even when it conflicts with common sense judgments about paradigm cases. To obey the standards of the orthodox metaphysics of causation is to hold an unreasonably high standard that unwisely excludes accounts that excel by all ordinary scientific criteria.

{ 14 }

The Psychology of Culpable Causation

Though causal culpability is metaphysically superfluous, it undoubtedly plays a prominent role in how we think about causation, including many of our explanatory practices. An adequate account of the metaphysics of causation ought to play a role in explaining why it is reasonable for humans to believe in culpable causes and why we have certain shared intuitions about culpability. Orthodox metaphysical accounts explain the reasonability of such beliefs by claiming in effect that these beliefs are true in the most literal sense. There are cause-effect relations out there in reality (in many cases holding between fairly localized singular events) as part of the world's metaphysical structure and people have a more or less accurate epistemic grasp of them. According to my account, belief in culpable causes is reasonable because there exist (metaphysically fundamental) terminance relations and (metaphysically derivative) prob-influence relations, and our intuitions about culpability serve as cognitive shortcuts for dealing with them.

In this chapter, I will construct a toy psychological theory whose primary purpose is to illustrate how my account of causation leads rather naturally to several heuristics for judging culpable causation. The toy theory shows how culpable causes help us learn about prob-influence along the lines of the discussion in §8.2. A secondary purpose of the toy theory is to complement my argument for locating culpable causation in the top conceptual layer of causation by demonstrating how many alleged problems in the metaphysics of causation dissolve once we acknowledge that a theory of culpable causation can be acceptable and informative and explanatory even if it has genuine conflicts and thus does not satisfy STRICT standards of adequacy. Once we reject that we should hold out for a complete and consistent systematization of cause-effect relations "out there in reality" that correspond to our folk conception of causation (or some moderately regimented version of it), many traditional puzzles about causation are easily resolved.

It is not my aim to provide anything remotely close to a full theory of the psychology of causation, nor even to provide a comprehensive theory of how people make judgments about culpable causes because that would be far too ambitious a topic. It would also distract from the main task of demonstrating that there is a reasonable link between my metaphysics of causation and the psychology of causation, broadly construed to include causal explanation. Furthermore, in order to keep this chapter as concise as possible, I have had to relegate some standardly discussed topics to an extended version of this chapter that I have made available.

Although I have attempted to construct the psychological theory in this chapter to accord with a wide range of stock intuitions about causation, it deserves to be called a toy theory for three reasons. First, it is a woefully simplistic theory that does not take into account the wide range of psychological data relevant to this topic and is only intended as a preliminary gesture.

Second, it does not produce any quantitative psychological predictions. For example, it does not provide enough structure to predict how much people's confidence in their judgments will change as they consider hypothetical situations that are ever more remote from ordinary experience. The toy theory does suggest some crude default predictions, but because I am unable to offer any principles that indicate where its predictions will be overridden by a more sophisticated treatment, there is no sure way to tell which failures of the default predictions are a result of its being based on an inherently defective scheme and which are merely the result of its being the toy theory it purports to be. So, whatever seeming success the toy theory has at explaining our common-sense intuitions about culpability should be weighed against the fact that it is not risking falsification with any bold predictions as a more serious theory would. (Also, I cannot address how the toy theory of culpability could be integrated with an account of the psychological mechanisms needed to implement assessments of culpability.)

Third, I am not pretending that the theory is free of counterexamples. On the contrary, one of my aims in discussing the toy theory is to illustrate a theory of causation that only meets RELAXED standards of adequacy. I will deliberately provide conflicting rules of thumb for identifying culpable causes in the technical sense of 'conflict' from §1.8. As foreshadowed in §1.10, my toy theory will not provide any formal rules sufficient to ameliorate these conflicts but will instead blithely delegate the conflict-resolution to my metaphysics of causation. In other words, whenever the rules of thumb I present for evaluating whether *C* is a culpable cause of *E* result in contradictory judgments in some realistic scenario, my theory declares that if you want consistency, you either (1) select one of the rules of thumb that is generating the consistency and stipulate that it is inapplicable to the scenario being considered, or (2) forgo talk of culpability in favor of contribution. You say it isn't clear whether *C* is a cause of *E* according to my theory? Fundamentally, all the contributors are partial causes of *E*, and there is always a definitive answer as to whether one fundamental event is a contributor to another. The more restrictive conception of singular cause that I have labeled 'culpable cause' is useful for epistemological purposes like causal explanation and discovering promotion relations, but these practices do not require STRICT consistency; a system of managed inconsistency is adequate.

Remember that because the purpose of the toy theory is to complement the metaphysics, its shortcomings do not undermine the metaphysical system provided in previous chapters. Psychological considerations could serve as evidence against a metaphysical account of causation only if the metaphysics were to make highly implausible the provision of a reasonable account of how humans could have the shared intuitions about causation that they have.

14.1 The Toy Theory of Culpable Causation

My metaphysics of causation says that (1) fundamentally, causation consists of terminants and contributors, which play the role of full and partial singular causes respectively;[27] and (2) we can abstract away from this kind of singular causation to get promotion relations, which adequately characterize general causation. If this is correct, our folk conception of singular causation among mundane events—culpable causation—is our imperfect way of grasping facts about terminance and promotion and the like.

Because one of the main reasons we have a notion of culpable cause is that it aids our discovery of promotion or prob-influence relations—a hypothesis I suggested in §8.2—we should expect this function to reveal itself in our judgments. It will turn out in §14.4 that there are discrepancies between what we would judge culpable if we cared only about whether that one particular instance of C affected the probability of that one particular instance of E and what we would judge culpable if we cared more about the *discovery* of prob-influence relations that apply to more general circumstances. When such discrepancies appear, according to my theory, we should expect our instinctive judgments concerning culpable causes to track the latter because such thinking would have greater practical utility.

A tension inherent in the idea of culpable causation is that it is a notion of *singular* causation that tries to incorporate features that essentially belong to *general* causation. On the one hand, it purports to apply to individual fragments of history, and, on the other hand, it privileges some contributors as more important to the occurrence of the effect than others. But the causal significance of each contributor in a single case ultimately derives from the fact that some kinds of events are *generally* good at bringing about other kinds of events. Culpability is what we get when we try to project onto individual fragments of history principles that govern general causation. Our implicit rules for assessing culpability are structured to mitigate the tension between the singular and general aspects of causation, but they do so imperfectly. Some of the implicit rules are easy to evaluate, but are less valuable as a guide to promotion relations. Others are harder to evaluate but provide a better guide to promotion relations. None of the rules carve nature at the joints. Our implicit conception of a culpable cause is a kludge that serves us well enough in practice, but whose implicit rules arguably do not systematize in a fully coherent way.

I think the core idea at the heart of culpability is this:

> An event is a ***culpable cause*** of E iff it successfully induces E.[28]

[27] Recall again that there are several important respects in which terminant relations do not match what we intuitively think of as causal, e.g., by being reflexive and not necessarily being asymmetric.

[28] This guiding principle is one variant of the hypothesis that singular causation can be adequately understood in terms of probability-raising processes. This should not be surprising because such theories are motivated primarily by the goal of incorporating (1) some sort of production or process or mechanism with (2) some sort of counterfactual dependence or difference-making or probability-raising. Because the metaphysics of causation I have presented represents

To begin the investigation of this guiding principle, I will first impose a simplifying assumption, second comment on 'induces', and third comment on 'successfully'.

First, in assessing causal culpability, the starting materials include (1) a sufficiently filled-in scenario, which is a possible fragment of history with some sort of laws governing its temporal evolution, and (2) a chosen occurrence in that scenario, the effect. The goal is to identify any happenings in that fragment of history that deserve to count as "one of the causes" of the effect. If I were unconcerned with overly cluttering the discussion with technicalities, I would make explicit that my discussion of culpable causation is compatible with the hypothesis that space-time is metaphysically derivative. But because the required terminology might be confusing, I will present this chapter (without loss of generality) under the assumption that some sort of space-time is the (fundamental) arena.

Second, I have introduced the term 'induce' to serve as a rough and ready psychological surrogate for 'promotion'. Because our native conception of culpable causation does not take into account the vast background that is usually required for promotion, it is best to avoid defining culpability exclusively in terms of promotion. We have at least some grasp of the idea that one event C can help make E occur. One could say that C-events have a tendency to result in E-events, C-events lead toward E-events happening, or C-events have a causal power to bring about E. In this chapter, 'induce' should be interpreted liberally enough to accommodate this variety of ways in which a cause can help make an effect come about.[29] Nevertheless, in order for me to connect the toy theory of culpable causes to my formally defined relation of promotion, it will facilitate communication if "C induced E" is primarily understood as "C raises the probability of E" which in turn can be related to promotion insofar as talk of plain coarse-grained events like C can be translated into the language of contrastive events.

When sorting through various candidate causes of an effect, we normally think of

fundamental causation along the lines of (1) and derivative causation along the lines of (2), the theory of culpability that complements the metaphysics should incorporate both aspects. There are some existing proposals along these lines, like Schaffer (58), but I do not know of any account that resembles the version presented in this chapter.

[29] I invite readers to interpret 'induces' liberally enough to include models of causal tendencies expressed in terms of forces or hastening or intentions. For example, there is a sizable literature in psychology based on the suggestion of Talmy (66) that many of our intuitions about causation can be effectively modeled in terms of our conception of force vectors. Wolff and Zettergren (69) report that a force-based approach successfully predicts a range of causal judgments regarding material objects. For example, if a motorboat is attempting to go away from a buoy but a strong wind blows it back until it hits the buoy, people will say the wind caused the boat to strike the buoy. Also see Wolff (70). The pronouncements of this "force dynamic" model of causation, I believe, overlap enough with the pronouncements one gets from a well-designed model based on difference-making in order to justify the following claim. If it is useful for a creature to possess the psychological faculties described by one of these two models—the "force dynamic" model or the difference-making model—it is useful for a creature to possess the psychological faculties described by the other. Similar comments apply to cases where someone hastens the occurrence of an effect that would have happened later without the action and to cases where someone acts intending for a certain effect to occur. The scenarios where these models disagree are important for debates in psychology, but I will not be concerned with their differences because the toy theory is only intended to establish a fairly reliable link between promotion and our assessments of culpability, not to insist that people's reasoning about causal tendencies must closely match probabilistic relations.

each candidate, c_i, under some not-too-convoluted coarse-grained description, C_i. For brevity, I will use the expression "c (as C)" to refer to the fine-grained event c under the coarse-grained description C. In order for it to be connected to the metaphysics, though, the event also needs to be thought of as a contrastive event, $\tilde{C}$, which comports with the observations of §4.8 that we often tend to use implicit contrasts when thinking of culpable causes. In all cases that we need to consider, the contrastive event is intended to be a contrastivization of the coarse-grained event having its background conditions filled in with a reasonable contextualization of C's actual environment at the same time as C. For brevity, I will use the shorthand "c (as C qua $\tilde{C}$)" to signify that c has been coarse-grained as C and contrastivized as $\tilde{C}$.

The practice of switching between coarse-grained and contrastive events applies to the effect as well. In order to keep the discussion in this chapter manageable, I will initially treat effects as plain coarse-grained events. In §4.8, I described how my account can handle contrastive effects as well, illustrated by the statement, "Adding a dash of salt causes the dish to be tasty rather than bland." Such contrastive effects can be accommodated by considering fixing relations rather than prob-influence relations. For example, in seeking the culpable causes of the dish being tasty rather than bland, we would ignore events like the presence of working kitchen equipment and the presence of groceries. These are promoters of the dish being tasty rather than not existing at all, but they are not promoters of the dish being tasty rather than bland. So, throughout the rest of this chapter keep in mind that my talk of promoting the effect E is meant to extend to contrastive effects and the events that fix them.

Third, as we proceed through the following discussion, I will progressively spell out four candidate interpretations for 'successfully' in the definition of 'culpable cause'. This will result in four distinct formulations of culpability. Each successive version builds on the previous one in order to match our instinctive identification of culpable causes better. I will first lay out the simplest version of culpability, culpability$_1$, to establish a basis for (1) clarifying how the effect and its potential causes are individuated, (2) specifying some parameters people tend to employ when judging promotion, and (3) exploring a preliminary guess at what it means for an instance of promotion to count as successful. Then, I will examine some deficiencies of culpability$_1$ in order to motivate an improved conception, culpability$_2$, which takes into account the contrastive character of causes and the fine-grained character of the effect. After explaining how culpability$_2$ addresses the problems with culpability$_1$, I will reveal some deficiencies culpability$_2$ has by virtue of its not taking into account anything that occurs temporally in between a candidate cause and the effect. Culpability$_3$ modifies culpability$_2$ by taking into account intermediate happenings, which allows it to be more discriminating by ruling out some candidate causes for failing to deliver their inducement successfully through an appropriate process. The final notion, culpability$_4$, extends culpability$_3$ by chaining together instances of culpability$_3$. I will then attempt to connect these last two technical notions to our intuitive conception of culpability, suggesting that we tend to vacillate between culpability$_3$ and culpability$_4$ depending on our explanatory purposes. Culpability$_1$ and culpability$_2$ merely serve as heuristic devices to help me communicate the content of the toy theory and to illustrate how it addresses standard examples in the philosophical literature on causation.

14.2 Culpability$_1$

Here is an initial refinement of the schematic definition of causal culpability:

> An actual event c (as C qua $\tilde{C}$) is ***culpable$_1$*** for an actual event e (as E) iff $\tilde{C}$ is a salient, significant promoter of E.

Culpability$_1$ captures the idea that culpability is successful promotion in the most naïve way possible. The cause occurred; it promoted the effect; the effect occurred.

14.2.1 SALIENCE

A ***salient promoter*** is a promoter people tend not to ignore as part of the causal background. In the psychology literature, the expression '***focal set***' refers to the set of contextually salient events that serve as candidate causes. There is a sizable literature on principles that determine which events are part of the focal set, and a more sophisticated account of culpability would presumably benefit from being integrated with a general psychological theory of focal sets, but that is far beyond the scope of this discussion. I will just mention a few issues that are particular to my toy theory.

The striking of a match counts as a salient promoter of its flame whereas the presence of oxygen does not, even though either one alone would not promote the flame in the absence of the other. What makes the striking stand out more than the oxygen has little to do with its role in nature and a lot to do with how we think of it. Reasons for conceiving of a promoter as worthy of special consideration include that it is the action of an intentional agent, that it is an unusual event, or that it deviates from what should be happening either in a moral sense or in the sense of an object performing its perceived function or in the sense of an object's deviating from its inertial path.[30] The implicit contrasts we use to select promoters play a large role in the process of identifying salient events. When an event takes place that is commonplace and either unchanging or in accordance with how things are supposed to be, we tend not to notice a contrast and therefore tend not to flag the event for further consideration. Most of the reason the presence of oxygen does not count as salient is that oxygen is almost always present at the Earth's surface and so we tend not to think of its absence as worth considering. The striking of a match counts as salient largely because it is an intentional action, involves a noticeable change, and is much rarer than other promoters like the presence of oxygen or the dryness of the match.

Some evidence exists that moral categories play a role in our selection of which events potentially count as causes, for example, Alicke Alicke (4), Knobe and Fraser (36), and Driver (12) (13). This would be surprising in a model of moral judgments where step one is to ascertain which events count as causally relevant without any appeal to

[30] See Maudlin (47) and the discussion of default and deviant states in Hall (23) and Hitchcock (33).

morality, and step two is to apply moral principles to assess those events for moral culpability. Although investigation of the role of morality in people's identification of culpable causes is in its infancy, the claim that our beliefs about morality play a role in whether some chosen event counts as a cause would not be surprising given my theory of causation. Because the concept of culpable cause is parasitic on the notion of promotion, culpable causes inherit the contrastivity of promotion. And, as noted in §4.8, the default contrasts people use in assessing causal promotion include what people believe is normal or what they believe should happen. We can think of "what should happen" as what typically happens, or as what will happen if things work as they are intended or designed to function, or as what the law or morality dictates. All these senses of 'what should happen' can play a role in identifying candidate causes. For example, when determining why a particular bridge collapsed, we tend to sift through events that differ from the norm in one of these senses. We might flag the existence of an unusually heavy load as a candidate cause just because it is atypical. Or we might flag the failure of a certain joint to maintain rigidity as a candidate cause because the purpose for which it was installed was to hold its beams rigidly together. Or we might flag the inspector's negligence because he was legally obligated to check the joints and morally obligated to make a good-faith effort. Actions people take in accordance with the law and morality are ceteris paribus less likely to be salient because routinely considering them would usually result in an unmanageably large number of candidate causes.

Another factor governing whether an event counts as salient is how broadly it is coarse-grained. The coarse-graining is often selected by some sort of default conception of an event, but we also have the ability to select a coarse-graining as salient in a more sophisticated manner. Imagine observing a person who is the subject of a psychological experiment. The subject attends to an unlit button on a panel; the button lights up with a green color; and the subject responds by pressing the button. It is natural to conceive of the situation as one where the lighting of the button caused the person to press it or where the lighting of the button as green caused the person to press it. One would not normally think of the cause as "the button lighting as either green or yellow" because there is no reason to suppose the button can light up as yellow or that a yellow light would induce the subject to press the button. However, if you are told the subject was instructed to press the button when and only when the light appeared as either green or yellow, and you see the button turn green and then the person pressing it, it would be reasonable for you to describe the cause as "the button lighting up as green or yellow." That description is appropriate because you know the most informative description of what is promoting the person to press the button is its lighting up as either green or yellow. It is reasonable to select this "green or yellow" contrastivization to inform one's selection of a salient candidate cause even though nothing in this particular case prevents one from accurately describing the cause more narrowly as "the button turning green." (Communication of the intended contrast also plays a role here.) This feature accords with my contention that our intuitions about culpability are often tuned in order to be useful for conveying information about promotion. Unlike Yablo's (72), (73) principle that "causes must be proportional to their effects," however, culpable causes in my account need not be coarse-grained in a maximally informative way.

There is also quite a bit of flexibility and lack of specificity not only in how we select some contrasts as the appropriate ones for defining the candidate causes but also in how broadly to coarse-grain the background conditions. The key idea motivating culpability$_1$ is that we have some conception of what it is for an event to be generally good at bringing about the general kind of effect that E represents. To cash out this idea appropriately, the background conditions implicit in the contrastive events need to be both broad enough and specific enough to capture an ordinary understanding of the general conditions under which events occur. For example, in identifying the culpable$_1$ causes for a campfire, we typically seek promoters of campfires by considering contrastive events that range over a wide range of earthly environments. But the relevant extent of the contrastive events does not extend to include conditions present in deep space or at the bottom of the sea. The appropriate degree of generality is something that the toy theory leaves as a rather flexible parameter.

14.2.2 IRREFLEXIVITY

Even though an event always determines or fixes itself, we generally judge that events do not cause themselves. This can be explained by noting that an event's self-determination or self-fixing is entirely trivial in the sense that it holds regardless of the laws and regardless of the character of the event. The triviality is pragmatically evident in the pointlessness of adopting the strategy to bring about E by bringing about (some contextualization of) E. Also, in presenting a causal explanation for E, it would be pointless to cite E as a cause because that would provide no new information. Because trivial fixing relations are always useless in practice, it makes sense for humans not to think of them as instances of causation at all. In general, to represent this pragmatic feature, we can simply declare that as a rule, no event is culpable for itself. It is conceivable that this rule might be overridden, perhaps to make a theological point, but it is reasonable to suppose it holds generally of mundane events.

14.2.3 ASYMMETRY

Because past-directed prob-influence is apparently always useless for the advancement of goals, it is reasonable for us to conceive of the past as settled and thus to think of events as not genuinely promoting past effects. If we instinctively think of events as not promoting past effects, it is reasonable for us not to count any events as culpable for previous happenings. This general rule can be overridden by prompting people with time travel stories or tales of magical past-affecting wands, and to the extent that people come to accept the possibility of such past-directed promotion—often because it is of a kind useful for advancement—they can come to override the default rule of thumb that events do not cause anything toward the past.

This explanation of the asymmetry of culpable causation in terms of the advancement asymmetry leaves open the possibility that a pair of simultaneous events can be culpable causes of each other. Because it is plausible that the actual laws obey the non-spatiality of terminance, as defined in §2.4.4, it is reasonable to guess that non-trivial simultaneous promotion does not exist. Alleged instances of simultaneous causation, e.g. Huemer and Kovitz (34), such as the causation existing between two nearly upright books that are tilted to prevent each other from falling, are not genuine cases of simultaneous causation. Every temporal stage of each book is a promoter of the other book remaining in place for the short-term future, but not a promoter of the other book being where it is at that very same instant. However, because humans typically select salient causes that are temporally extended, it would be understandable for people to ignore the subtle details of timing and speak of simultaneous causation in such cases. That said, genuine simultaneous causation (in the sense of space-like contribution) is certainly a coherent possibility, and there is no inconsistency in the hypothesis that two space-like separated events could be culpable causes of each other.

In §4.12, I noted the existence of non-local partial influence. Although I believe that provides a legitimate, albeit esoteric, sense in which two simultaneous events can non-trivially promote each other, I believe it is far enough removed from the way people ordinarily think of causation, to disregard it when theorizing about the psychology of causation. People might occasionally employ reasoning that corresponds to pseudo-backtracking connections, but because non-local partial influence is only exploitable in the way described in §6.4, it makes sense for people, upon a modest amount of reflection, to interpret non-local partial influence as not being genuinely causal, even though according to my theory it really is.

14.2.4 SIGNIFICANT PROMOTION

When an event C increases the probability of the effect from nearly zero to some appreciably large value and the effect occurs, we tend to think of C as a culpable cause, barring some reason to think otherwise. But in many cases, the promotion is not significant enough in magnitude to warrant our assigning it culpability for the effect. Judgments of significance are guided in part by the absolute amount by which the probability of E is increased, but there is an asymmetry in how we treat probability raising when it involves unlikely events compared to when it involves likely events. For example, if C and E both occur and E had a 99.9999% chance of occurring in the presence of C but would have had a 99% chance of occurring in the absence of C, then people will be less likely to classify C as a cause than they would if C raised the chance of E from 0.0001% to 1% despite the same increase in the absolute magnitude of probability. This difference in judgment is understandable in terms of either of the two following psychological rules. The first is that we reckon probability-raising at least partly in terms of ratios, not absolute increases. When the contrast probability is lower, the degree of promotion will be a higher factor; 1% is ten thousand times greater than 0.0001% whereas 99.9999% is barely greater than 99%. The second possible psychological rule is that we think of culpability as something which itself is susceptible to chance. The subject may know that C increased the probability of E from 99% to 99.9999% but recognize that E probably would have happened anyway and thus judge that C only had a relatively small chance, maybe around 1%, of being something that made a difference to *E*'s occurrence.

Another aspect of judging whether the promotion is significant enough occurs when the resulting chance of the promoted effect is still small. If *C* raises the probability of *E* from 10^{-100} to 0.01, and there are no other candidate promoters, and *E* occurs, then we tend to identify *C* as a cause of *E*. Other cases, though, are less clear. Suppose the causal background is such that the event *E* has a 10^{-29} chance of occurring without any salient cause. If some *C* raises the probability of *E* from 10^{-29} to 10^{-20}, it has increased the chance of *E* a billion-fold but only raised it to a minuscule level. In such cases where *C* occurs, followed by *E*, it can be unclear whether we should attribute *E*'s occurrence to *C*.

In addressing this question, a potential ambiguity in the ordinary notion of cause is exposed, which I previously mentioned in §1.10 and §4.5 and the introduction to chapter 8. Sometimes we think of a cause of *E* under the description 'one of the causes of *E*' and at other times under the description 'something that caused *E*'. These two descriptions do not always pick out the same events. When Lori buys a lottery ticket and wins, we ought to say her purchase of the lottery ticket was one of the causes of her winning, but we also ought to say her purchase did not cause her to win, presumably because it did not raise the probability to a high enough level. So, her purchase was one of the causes of her winning but was not something that caused her to win. I defined 'culpable cause' to be equivalent to the 'one of the causes' reading and not the 'something that caused' reading. So, Lori's purchase was culpable for her winning. (One might think that the 'something that caused *E*' reading of '*C* is a cause of *E*' tries to capture the idea that *C* is the dominant cause of *E* among all the culpable causes, but I think a better way to put it is that *C* is the dominant cause among all those culpable causes that occur at the same time as *C*. For, if someone topples a row of twenty dominoes and the last falling domino rings a bell, *E*, it is correct to say of each fallen domino that it is something that caused *E*.) The ambiguity in the clause 'is a cause of' will recur in other examples in this chapter.

14.2.5 CAUSAL GROUPING PRINCIPLES

Promotion is defined in terms of a contrast, and the default contrast people tend to use for a candidate cause (that they implicitly conceive as a localized event *C*) is to hold the actual background conditions fixed and replace what is going on at *C* with some contextually appropriate happenings that do not instantiate *C*. However, this default rule for selecting contrasts is only a crude approximation of how people think. Sometimes we have other heuristics for selecting a background that result in alternative contrasts. An important example where people may override the default rule for selecting contrasts is the case of overdetermination. In this section, I will discuss overdetermination and the closely related concept of joint causation.

In chapter 2, the concept of overdetermination was discussed with regard to multiple fine-grained events determining the same event. However, there is an altogether different notion of overdetermination that pertains to culpable causation. Regarding

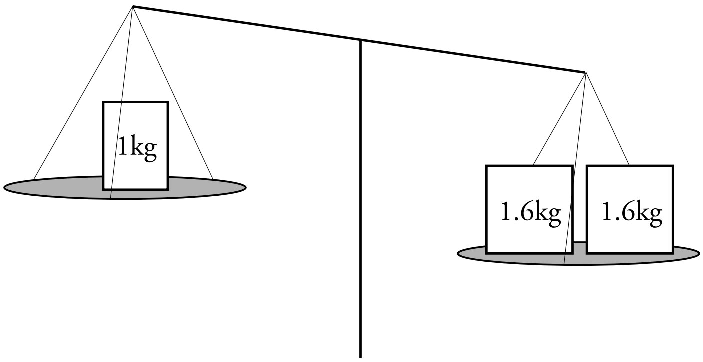

FIGURE 14.1 *Each 1.6 kg mass (in the absence of the other) promotes the balance being tipped right. Neither mass (in the presence of the other) promotes the balance being tipped right.*

culpability, ***overdetermination*** occurs when multiple distinct events are culpable as a group for some effect, and also individually. I will define overdetermination only for the simplest example where there are two salient events.

> Two distinct existing events c_1 (as C_1) and c_2 (as C_2) are overdetermining culpable$_1$ causes of e (as E) if and only if all of the following hold:
>
> 1. $p_{\overline{C_1 \& C_2}}(E) \gg p_{\overline{\neg C_1 \& \neg C_2}}(E)$
> 2. $\neg\left(p_{\overline{C_1 \& C_2}}(E) \gg p_{\overline{C_1 \& \neg C_2}}(E)\right)$
> 3. $\neg\left(p_{\overline{C_1 \& C_2}}(E) \gg p_{\overline{\neg C_1 \& C_2}}(E)\right)$
> 4. $p_{\overline{C_1 \& \neg C_2}}(E) \gg p_{\overline{\neg C_1 \& \neg C_2}}(E)$
> 5. $p_{\overline{\neg C_1 \& C_2}}(E) \gg p_{\overline{\neg C_1 \& \neg C_2}}(E)$

where '$\gg$' means 'significantly greater than', as discussed in §*14.2.4*, and the obvious contextualizations are employed.

For illustration, consider a balance with a 1kg weight on the left. When two 1.6 kg masses are placed on the right side, the balance tips to the right as depicted in Fig. 14.1. Let c_1 be the placing of one 1.6 kg mass on the right side, and let c_2 be the placing of the other 1.6 kg mass on the right side. The two events are clearly culpable$_1$ together because had the pair of masses not been placed on the balance, the balance would not have tipped to the right. But what about the culpability$_1$ of each individual event? Using the default background condition where we hold the presence of the other mass fixed, we get the result that each event by itself is not culpable$_1$ for the balance tipping. After all, the other mass would still have been placed and the balance would thus have tipped to the right. However, it is also possible to construe c_1 and c_2 under an alternative contrastivization where we think of the other mass as being absent and then evaluate whether the event is culpable$_1$ for the balance tipping. Under that construal, each event is successful at promoting the tipping of the balance because each one has enough mass by itself to tip the balance. When events are culpable$_1$ together and they are not individually culpable$_1$ using the default contrastivization (drawn from the way things are actually laid out) but they are successful promoters using the non-standard contrastivization where the presence of the other event is written out of the background conditions, then the effect is overdetermined by the two events.

Joint causation occurs when multiple candidate causes are culpable as a group but not individually with respect to a contrast where neither of them is present. I will define joint causation only for the simplest example.

> Two distinct existing events c_1 as C_1 and c_2 as C_2 are joint culpable$_1$ causes of e as E if and only if all of the following hold:
>
> 1. $p_{\overline{C_1\&C_2}}(E) \gg p_{\overline{\neg C_1\&\neg C_2}}(E)$
> 2. $p_{\overline{C_1\&C_2}}(E) \gg p_{\overline{C_1\&\neg C_2}}(E)$
> 3. $p_{\overline{C_1\&C_2}}(E) \gg p_{\overline{\neg C_1\&C_2}}(E)$
> 4. $\neg\left(p_{\overline{C_1\&\neg C_2}}(E) \gg p_{\overline{\neg C_1\&\neg C_2}}(E)\right)$
> 5. $\neg\left(p_{\overline{\neg C_1\&C_2}}(E) \gg p_{\overline{\neg C_1\&\neg C_2}}(E)\right)$

where the obvious contextualizations are employed.

For illustration, consider the balance with just a 1 kg weight on the left. When two 0.7 kg masses are placed on the right side, the balance tips to the right as depicted in Fig. 14.2. Let c_1 be the placing of one 0.7 kg mass on the right side, and let c_2 be the placing of the other 0.7 kg mass on the right side. The two masses together are culpable$_1$ for the balance tipping to the right because had they not been placed on the balance, it wouldn't have tipped to the right. But what about the culpability$_1$ of each one individually? Using the default background conditions where we hold the presence of the other mass fixed, we get the result that each event by itself is culpable$_1$ for the balance tipping. After all, in the presence of the other mass, each event would have promoted the balance tipping. However, it is also possible to construe c_1 and c_2 under an alternative contrastivization where we think of the other mass as being absent and then evaluate whether the event is culpable$_1$ for the balance tipping. Under that construal, each event is unsuccessful at promoting the tipping of the balance because a single 0.7 kg mass is not enough to tip the balance. When events are culpable$_1$ together and they are individually culpable$_1$ using the default contrastivization (drawn from the way things are actually laid out) but they are not successful promoters using the non-standard contrastivization where the presence of the other event is written out of the background conditions, then the effect is jointly caused by the two events.

My reason for mentioning these grouping principles is that people are capable of judging culpability based on different ways of grouping events, and concepts like overdetermination and joint causation provide a richer picture of the underlying promotion relations, especially when the promotion relations occur in complex combinations. For example, one could have a situation where a group of five events is together culpable$_1$ for an effect E. To evaluate culpability$_1$ for each individual event, one would consider the default background conditions where the occurrence of the other four events is held fixed. But one could also consider other non-standard background con-

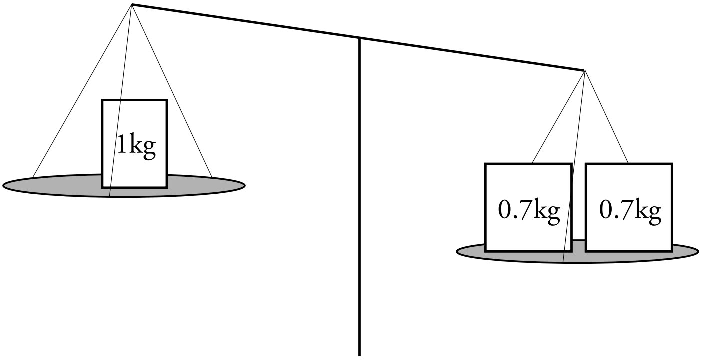

FIGURE 14.2 *Each 0.7kg mass (in the presence of the other) promotes the balance being tipped right. Neither mass (in the absence of the other) promotes the balance being tipped right.*

ditions, and there are many combinations to consider. One could evaluate whether C_2 promotes E in the presence of C_1 and C_3 but not C_4 and C_5, or one could consider whether the C_2 and C_3 together promote E in the presence of C_5 and the absence of C_1 or C_4. There are many ways to combine events into groups, and because the patterns of promotion might be quite complicated, it can become unclear how to assess the culpability of the individual components once we stray from the default contrastivization (as we sometimes do).

In this section, I have identified several factors that play a role in settling whether a candidate cause of E is culpable$_1$ for E. An existing event is a culpable$_1$ cause of some effect E if and only if (1) it is a member of the focal set of contextually relevant events, (2) it is not E itself, (3) it temporally precedes E, and (4) it is a significant promoter of E. This characterization should be understood in light of the qualifications and emendations I have suggested in this section.

14.3 Shortcomings of Culpability$_1$

Culpability$_1$ measures the success of $\tilde{C}$'s significant promotion of E in the crudest way possible. The promotion is successful if and only if E occurs. In this section, I will examine some deficiencies of this measure of success by providing several examples where culpability$_1$ fails to match some pre-theoretical judgments concerning culpable causation. I will respond to these faults in the next section by defining an improved concept, culpability$_2$.

14.3.1 PRECISE CHARACTER OF THE EFFECT

Consider a fragment of history with two cannon-like machines that launch paint balls toward a single canvas mounted on a wall. The machine on the left is able to hit the canvas with 99% accuracy and selects its paint balls from a random assortment of one hundred different hues, not including periwinkle. The machine on the right is able to hit the canvas with 1% accuracy and only uses periwinkle-colored paint balls. The machines are fired simultaneously once and a single paint splat forms on the canvas, which happens to be periwinkle in color. Let the fine-grained effect e be the full state (five seconds after the machines are fired) of the canvas and its immediate environment, including any parts of the wall within a few meters of the canvas. Let C_l and C_r be the firing of the left and right machines respectively, and let E be the event of the canvas having paint on it five seconds after the firing. Which of the machines were culpable for E? Our intuitive judgment selects C_r and not C_l by virtue of the fact that the right machine is the only one capable of making a periwinkle splat. But C_l is culpable$_1$ for E because C_l and E occurred and C_l raised the probability of E significantly over what it would have been had the right machine fired alone. Thus, culpability$_1$ does not match our instinctive identification of the culpable causes.

14.3.2 OVERLAPPING CAUSATION

Now suppose the left machine is aimed slightly to the left of the canvas so that when it splatters paint onto the canvas, it also splatters paint to the left of the canvas and it never splatters to the right. Suppose the right machine is aimed so that it splatters to the right when it hits the canvas and never splatters to the left. Suppose that both use green paint balls and that e instantiates a splattering of paint onto the canvas and onto the wall to the right of the canvas. Which machine caused the canvas to acquire paint? We tend to select C_r and not C_l. One good reason is that the right machine is the only one capable of making a splat that spreads to the right of the canvas. But C_l is culpable$_1$ for E because C_l and E occurred and C_l raised the probability of E significantly over what it would have been had the right machine fired alone. Thus, culpability$_1$ does not match our instinctive identification of the culpable causes.[31]

14.3.3 PROBABILITY-LOWERING CAUSES

Suppose as before that the left machine firing alone is 99% accurate and the right machine is 1% accurate, but now introduce an interaction between C_l and C_r so that if they fire simultaneously, the accuracy of the left machine drops to 1%. Suppose the precise event e that occurs is a splattering of green paint onto the canvas and on the wall to the right of the canvas but not to the left. Which machine caused the canvas to acquire paint? We tend to select C_r and not C_l, again because of the paint to the right of the canvas. But C_r is not culpable$_1$ for E because C_r lowered the probability of

[31] See also, Schaffer (55).

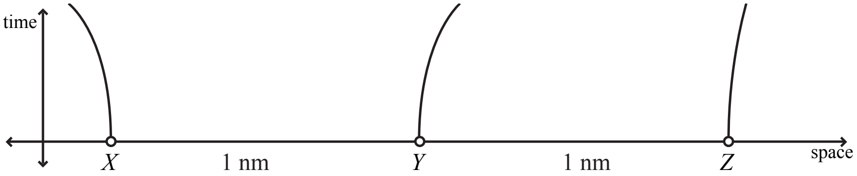

FIGURE 14.3 *Z's charge, but not Y's, contributes to the motion of X.*

E from 99% to approximately 2%. Thus, culpability$_1$ does not match our instinctive identification of the culpable causes.

14.3.4 TRUMPING

Consider the following model of fundamental physics. There exist three massive corpuscles, *X*, *Y*, and *Z*, in a Galilean space-time. There are two kinds of fundamental charge, weak and strong. *X* has +1 unit of weak and +1 unit of strong charge. *Y* has +1 unit of weak charge. *Z* has +1 unit of strong charge. The fundamental laws of this toy physics dictate a default rule that corpuscles interact via a classical Coulomb force law with respect to their weak charge properties and a separate Coulomb force law with respect to their strong charge properties. The default law governing the corpuscles is such that the two different Coulomb force terms add. So the impressed force on corpuscle *X* would normally be

$$F_X = k_w \frac{w_X w_Y}{(r_{XY})^2} + k_s \frac{s_X s_Z}{(r_{XZ})^2},$$

where r_{ij} is the distance between corpuscles *i* and *j*, k_w and k_s are constants, and w_i and s_i are the weak and strong charges of corpuscle *i* respectively. The direction of the force on *X* is determined by the standard principle that like charges repel and opposites attract.

However, suppose in this model that there is a special law dictating that whenever a corpuscle with a strong charge is within 5 nanometers of another strong charge, it is unaffected by any weak charges. In such situations, we say that the weak interaction is trumped by the strong interaction (56). If the three corpuscles are arranged as depicted in Fig. 14.3, the impressed force on *X* is just

$$F_X = k_s \frac{s_X s_Z}{(r_{XZ})^2} = k_s \frac{(+1)(+1)}{(2\,\mathrm{nm})^2}$$

directed to the left. Suppose that the corpuscles in Fig. 14.3 just happen to be in a special configuration where their motion exactly coincides with what their motion would have been without the special trumping law. In that case, an examination merely of the material layout of the fragment of history does not reveal whether the trumping law is operative. We can only infer that the weak charge is not contributing to *X*'s motion because of evidence that the trumping law holds generally.

To see the problem this case presents for culpability$_1$, consider the contrastive event, $\tilde{C}_Y$, that holds fixed the presence of *X* and represents the existence of *Y* at a distance of roughly 1 nanometer from X rather than a non-existence of Y. The rest of the background conditions of $\tilde{C}_Y$ are defined to include some probability of there being strongly charged particles nearby and some probability of there being no strongly charged particles nearby. It follows from the fundamental laws that $\tilde{C}_Y$ promotes the acceleration of X toward the left (due to the elements of $\tilde{C}_Y$ with no strongly charged particles around). For the sake of argument, suppose that the degree of promotion is significant enough for C_Y to count as a candidate cause. Then, C_Y is culpable$_1$ for X's particular leftward acceleration. However, considered reflection on the precise actual instance and the trumping law should lead people to judge that the presence of the Y particle is not culpable for X's motion because its impact on X is trumped by Z's presence.

14.4 Culpability$_2$

In an attempt to reduce the mismatch between culpability$_1$ and our instinctive judgments of culpability, we can define an improved successor concept, culpability$_2$. A metaphorical way to think about what makes culpability$_2$ essentially different from culpability$_1$ is that culpability$_1$ is the notion we get when we judge the success of a candidate cause merely by whether reality meets the goal we impose from the outside—our choice of how to coarse-grain the effect—whereas culpability$_2$ is the notion we get when we judge the success of a candidate cause in terms of whether the outcome it ended up inducing—an achievement defined in terms of a contrastive effect—also happens to promote the goal we have imposed from the outside.

To unpack this idea, let us first review how culpability$_1$ attempts to approximate our intuitive notion of culpability. The starting point for evaluating culpability is a fine-grained event, e, coarse-grained as E, which serves as the effect whose causes we seek. To find its causes, we look for events that induce E. An existing coarse-grained C induces E when some salient contrastivization $\tilde{C}$ of C significantly promotes E. Then, culpability$_1$ counts an instantiated event's inducement of E as successful if and only if E occurs.

The guiding idea behind culpability$_2$ is to measure successful inducement in terms of what results $\tilde{C}$ was "trying" to promote and how its attempt fared. Let us first define R to be the region occupied by our chosen effect E as well as a fair amount of its surrounding environment at the time t of E's occurrence. Second, $\tilde{G} \equiv (\overline{G}, \overline{\neg G})$ is defined to be the contrastive event occupying R that is fixed by $\tilde{C}$.[32] Then, we can think of $\tilde{G}$ in terms of its prominent foreground and background.[33] What $\tilde{C}$ was "trying" to promote is $\overline{G}$ rather than $\overline{\neg G}$, localized in $\tilde{G}$'s prominent foreground.

[32] This fixing of a contrastive event was defined in §3.7. For simplicity, I am assuming we are not dealing with cases of past-directed time travel or any other conditions that would allow $\tilde{C}$ to fix more than one contrastive event for R.

[33] Recall from §7.3.1 that the prominent foreground of $\tilde{G}$ is the region of the arena where $\overline{G}$ and $\overline{\neg G}$ differ significantly, and the prominent background is the complementary region where they do not differ significantly.

For an illustration, consider the case of overlapping causation from §*14.3.2* where the left machine is 99% accurate and can hit to the left of the canvas, and the right machine is 1% accurate and can hit to the right of the canvas, and the machines do not significantly interact with each other. The chosen actual effect e is a splattering of paint on the canvas and to the right of the canvas but not to the left. We stipulate our coarse-grained effect of interest, E, to be the existence of some paint on the canvas. Construed as a contrastive event, the firing of the right machine is $\tilde{C}_r \equiv (\overline{C_r}, \overline{\neg C_r})$, which fixes $\tilde{G} \equiv (\overline{G}, \overline{\neg G})$ for region R. In this example, $\overline{G}$ and $\overline{\neg G}$ are very nearly alike except that $\overline{G}$ is far more likely than $\overline{\neg G}$ to instantiate a paint splatter on the right side of the wall. Because $\overline{C_r}$ and $\overline{\neg C_r}$ very nearly agree on the likely motion of the left paint ball, everything that happens with the left paint ball is excluded from what is (trivially) promoted by (the prominent foreground of) $\tilde{G}$. What $\tilde{C}_r$ is "trying" to promote is not E, paint on the canvas, but paint being somewhere on the canvas and/or on the wall to its right rather than no paint in that region.

The next step in assessing culpability$_2$ is to characterize how $\tilde{C}_r$'s attempt at promotion "fared" in terms of a contrastive effect occupying R that is more finely-grained in order to account for what actually occurred (fundamentally) in R. Let us say that g_e is the unique full fundamental event occupying R. For concreteness let us suppose that g_e instantiates one splotch of paint on and to the right of the canvas, and another splotch of paint far to the left of the canvas. It also instantiates the wall, the lighting, the sounds, and other details in the room. We then slightly coarse-grain g_e as G_E to circumvent the problem that often all fundamental events equally have zero probability even though some are far more likely than others in the more intuitive sense captured by slightly coarse-graining them with some reasonable probability distribution.

We then construct a new event, $\overline{E_1}$, by using $\overline{G}$ as a starting point and removing all the members of $\overline{G}$ that are not members of G_E, renormalizing the resulting probability distribution to make $\overline{E_1}$ a well-defined contextualized event. The role of $\overline{E_1}$ is to represent in a slightly fuzzed way the portion of what $\overline{C_r}$ was "trying to make happen" that actually occurred at t. A reasonable choice of coarse-graining should fuzz $\overline{E_1}$ enough to eliminate stray traces of dust and similar fine details but not enough to eliminate the detailed pattern of the paint splatter or the brightness of the light on the wall or any audible sounds in the room.

We are in the process of constructing a contrastive event to represent the effect, and $\overline{E_1}$ serves as the protrast, the first member of the ordered pair of contextualized events. We now want to construct another contextualized event, $\overline{E_2}$, to serve as the contrast. In doing so, there are two main considerations we should attend to. The first is that we should restrict $\overline{E_2}$'s members to those that are already in $\overline{\neg G}$ so that we properly respect what $\overline{\neg C_r}$ was "trying to make happen." The second is that we should eliminate members of $\overline{\neg G}$ that would result in spurious identifications of promotion. I will return shortly to the question of how this is to be accomplished, but once the appropriate members have been removed and the probability distribution renormalized, we will have our sought-after contrast, $\overline{E_2}$.

The final step is to pair these two contextualized events together to form $\tilde{E} \equiv (\overline{E_1}, \overline{E_2})$, which represents everything $\tilde{C}_r$ successfully promoted for region R. To evaluate whether C_r successfully induced some chosen E, we now only need to consider the degree to which $\tilde{E}$ (trivially) prob-influences E. We can always read off of $\tilde{E}$ the degree of prob-influence for any plain coarse-grained event E in R as follows. The degree to which $\tilde{E}$ prob-influences E is equal to the proportion of $\overline{E_1}$'s members that instantiate a member of E (in the correct region) minus the proportion of $\overline{E_2}$'s members that instantiate a member of E (in the correct region). The degree to which $\tilde{E}$ prob-influences E is by construction equal to the degree to which $\tilde{C}_r$ successfully promoted E and thus is equal to the degree to which C_r successfully induced E. If this degree of successful inducement is significantly positive, then C_r counts as a successful inducer of E.

In order to fill in the gap left in this procedure—removing appropriate members from $\overline{E_2}$—there are at least two guiding principles we should apply. The first principle involves removing aspects of $\overline{E_2}$ that are not of the right kind to be prob-influenced by $\tilde{C}_r$. The second principle involves removing aspects of $\overline{E_2}$ that can be attributed to other independent causes. We remove an aspect of $\overline{E_2}$ by stripping out members of $\overline{E_2}$ to equalize the probabilities that $\overline{E_1}$ and $\overline{E_2}$ fix for that aspect. I will illustrate both principles with examples.

The first way to tell whether some aspect of $\overline{E_2}$ should be removed is to examine what kinds[34] of effects $\tilde{C}_r$ promotes for region R. Suppose for example that a particular location on the canvas and wall is lit by several spotlights that flicker on and off every now and then with $\tilde{C}_r$ not prob-influencing anything about the lights. Because $\tilde{G}$ does not prob-influence anything regarding amount of light striking the wall or canvas, we should adjust E_2 (to match $\overline{E_1}$ with regard to its pattern of luminosity) so that $\tilde{E}$ does not prob-influence the amount of light striking the wall. In this way, we render C_r not culpable for the canvas being lit the amount that it is, for the wall being at room temperature, for the existence of a roach at a particular location on the floor, and so on.

A special case of this principle involves transferring the prominent background of $\tilde{G}$ to $\tilde{E}$, as can be seen in our current example of overlapping causation where $\overline{E_1}$ fixes a very high probability for the particular pattern of paint on the wall to the left side of the canvas (which came from the left machine). Unamended, $\overline{E_2}$ would fix a very low probability for any particular splotch of paint because $\overline{\neg C_r}$ leaves open the full range of possibilities for where the left machine's paint can land. Unamended, $\tilde{E}$ would thus count as successfully promoting the splotch of paint on the left, which disagrees with our judgment that the firing of the right machine was not one of the causes of the left machine missing the canvas. To resolve this problem, it is reasonable to strip out members of $\overline{E_2}$ to make the prominent background of $\tilde{E}$ match the prominent background of $\tilde{G}$. The prominent foreground of $\tilde{G}$ is the subregion of R that includes everywhere the right paint ball could have landed, and its prominent background is everywhere else, including the actual location of the paint splotch on the left. The technical implementation of the solution is to discard any members of $\overline{E_2}$ that disagree with G_E in the prominent background of $\tilde{G}$, and then renormalize

[34] The relevant kinds here exclude any kinds that are too difficult for people to cognize.

its probability distribution. By doing so, the prominent background of $\tilde{E}$ will be the same region as the prominent background of $\tilde{G}$.

The second way to tell whether some aspect of $\overline{E_2}$ should be removed is to infer that this aspect is already attributable to some alternative cause that is independent (in the sense of not being significantly prob-influenced by) the candidate cause. When we have good grounds for attributing an aspect of the effect to another cause by virtue of some signature detail in what the alternative promotes, it should be removed from what $\tilde{E}$ promotes. When we do not have good enough grounds for attributing it to an alternative cause, then the culpability of C_r is not ruled out.[35]

For example, imagine a scenario where both green paint balls have landed on the canvas and overlap somewhat. $\overline{E_1}$ fixes a probability of one for the particular pair of paint splotches on the canvas, and without being further amended, $\overline{E_2}$ fixes a very low probability for paint being exactly at the location where the left machine's splotch of paint actually ended up. Consequently, $\overline{C_r}$ successfully promotes the left machine's paint hitting the target, which is the incorrect judgment. To remedy this situation, we should try to identify which aspects of $\overline{E_1}$ cannot be properly attributed to C_r (or are better thought of as attributable to causes other than C_r) and to conditionalize $\overline{E_2}$ accordingly to eliminate its promotion of those aspects. In our current example of the two overlapping splotches on the canvas, we can make a judgment as to which part of the paint pattern is attributable to the left machine's firing and discard from $\overline{E_2}$ any members that do not instantiate this part of the paint pattern. This alteration makes $\overline{E_2}$ agree with $\overline{E_1}$ as to the location of the left paint splotch and thus ensures that $\tilde{E}$ prob-influences the existence of the splotch from the left machine to degree zero, rendering C_r as not successfully inducing the left machine's splotch of paint landing where it did.

We can summarize this sketched procedure in terms of an semi-formal definition for culpability$_2$.

> An actual event c (as C qua $\tilde{C}$) is ***culpable*$_2$** for an actual event e (as E) iff a region R (including and surrounding e) has a contrastive effect $\tilde{E}$ imposed on it that significantly promotes E.

The imposed $\tilde{E}$ is generated by taking what $\tilde{C}$ fixes for R, conditionalizing its protrast with a slight coarse-graining of the full fundamental event occupying R, and adjusting its contrast in parallel in light of what $\tilde{C}$ and other independent salient events induce or promote for R.

Because this is a toy theory, I am forced to leave underspecified the precise implementation of the procedure for constructing the contrastive effect. For example, I cannot say whether some kinds of aspects are more salient for the purpose of stripping out aspects from $\tilde{E}$. In any case, one notable deficiency of the method described in this

[35] One could at this stage incorporate additional considerations related to causal grouping—overdetermination and joint causation—but unfortunately I have had to abbreviate this presentation of the toy theory.

section is that it does not work nearly as well when the contrast in the candidate cause is likely to interact with stuff in the environment and leave traces in the fine-grained effect. Unfortunately, I will have to forego how to refine the method further.

It is a good exercise at this stage to consider other cited shortcomings of culpability$_1$ in order to see how culpability$_2$ helps secure better agreement with our pre-theoretical judgments of culpable causation. In the example from §*14.3.1*, the right machine fires periwinkle paint balls and the left fires some other color. The firing of the left machine will not count as a successful inducer of E. Although $\tilde{C}_r$ promoted paint on the canvas, it did not promote the existence of periwinkle paint on the canvas, and yet periwinkle paint is the only color of paint on the canvas and thus was the only color of paint that was represented in the constructed $\tilde{E}$.

In the example from §*14.3.3*, the firing of the right machine lowers the probability of paint on the canvas, yet we still think of it as culpable for the existence of paint on the canvas. We can now make sense of this judgment. Even though $\tilde{C}_r$ makes paint on the canvas much less likely, it significantly promoted the probability of the precise pattern that happened to appear. C_r was successful at inducing the more finely grained effect and so was culpable$_2$ for E.

Given my previous suggestion that the reason we have a notion of culpability is that it allows us to more quickly infer promotion relations, it is worth considering why it would benefit us to have intuitions that match culpability$_2$ rather than culpability$_1$. I certainly do not think our psychological mechanisms for attributing singular causation implement the precise form of my exegesis of culpability$_2$. However, it is a plausible hypothesis that we reckon culpable causes by scouring the evidence included in the environment of the fine-grained effect and piecing together which candidate causes were responsible for which aspects by attributing each chosen aspect to a candidate cause (or candidate group of causes) when it is the only candidate that could have promoted that aspect.

In many circumstances, culpability$_2$ is not too much harder to assess, and it is often much more responsive to the observed evidence than culpability$_1$. As discussed in §8.2, recognizing which of the two machines is culpable for the effect often allows one to make good estimates about how much each machine individually promotes the effect. In circumstances where the paint-ball-firing machines barely influence one another, one can easily gather statistics on how often the splotch of paint spreads significantly toward the right and thereby infer the fraction of times that a C_r event was culpable for the existence of paint on the canvas, E. That, in turn, provides a good estimate for how likely the machine on the right would place paint on the canvas when operated alone.

14.5 Shortcomings of Culpability$_2$

The primary shortcoming of culpability$_2$ is that we sometimes rule out a candidate cause because it does not successfully deliver its inducement to the effect through an appropriate process. In technical terms, culpability$_2$ does not properly account for 'fizzling', a term from Schaffer (55).

Intuitively speaking, ***fizzling*** occurs when a process is "heading toward" bringing about E but reaches a stage where it is no longer bringing about E. Framed within the context of the toy theory, fizzling can be defined using the following procedure. First, assume that there is some actual event c (as C qua $\tilde{C}$) that promotes some E. There is no need to assume that E is instantiated. Second, we can consider any region R that is intermediate between $\tilde{C}$ and E, typically a region that lasts only for a moment. Second, let i be the actual full event occupying all of R. Third, construct a contrastive effect $\tilde{I}$ for the region R employing the same procedure used to evaluate culpability$_2$. Fourth, check whether $\tilde{I}$ significantly promotes E. If it does not, i counts as a fizzle with respect to E. If there is an actual event identified by this four-step process that counts as a fizzle, the process leading from c (as C qua $\tilde{C}$) to the promoted E counts as having fizzled.

A good example of fizzling occurs when a fuse is burning at time $t = 0$ and is "going to" launch a rocket at $t = 2$, and that nothing else of interest is going on. The default contrast built into $\tilde{C}$ at $t = 0$ is the fuse just lying there unlit with nothing in the background environment that would suggest that it could become lit in the near future. Suppose that shortly before time $t = 1$, the fuse burns out prematurely. In this case, the full event i at time $t = 1$ instantiates a burned out fuse, and the contrastive effect $\tilde{I}$ represents a short non-burning fuse rather than a long unlit fuse. Because this $\tilde{I}$ does not significantly promote the later rocket launch, i counts as a fizzle.

Let us now consider several examples where our knowledge of intermediate events motivates rejecting a candidate cause. These will illustrate how culpability$_2$ counts as a defective approximation of our instinctive concept of culpability.

14.5.1 SAVED FIZZLES

A ***saved fizzle*** is when there is some c (as C qua $\tilde{C}$) promoting some E, the process leading from c to E fizzles, and yet E occurs anyway. A simple case of a saved fizzle is when a lit fuse that is on its way toward launching a rocket spontaneously burns out for a while and then spontaneously becomes lit again and leads to the launching of the rocket. The spontaneous event here can be conceived as a highly improbable event that does not occur by virtue of any recognizable previous event but results from some fundamental or derivative chanciness. Intuitively, the initial lighting of the rocket was not one of the culpable causes of the rocket's launching, but it is culpable$_2$ for the launching because what actually occurred is very nearly what would have occurred had the fuse not burned out.

The next two subsections discuss cases of preëmption, that are also special cases of saved fizzles.

14.5.2 EARLY CUTTING PREËMPTION

Preëmption occurs when some event is culpable for a fizzle. ***Early cutting preëmption*** occurs when the caused fizzle precedes the induced effect. For illustration, consider the pair of machines that fire paint balls, but also introduce a shield that can spring into place and absorb one of the two paint balls without leaving any noticeable trace of which ball it absorbed. Suppose both machines are placed very close together and aimed so as to fire green paint balls in very nearly the same direction toward the middle of the canvas, so that the pattern of paint each would likely produce is the same. The machines are fired at the same time, but the left ball by chance happens to be absorbed by the shield, and the right paint ball lands on the canvas. Which machine caused the canvas to acquire paint? We tend to select C_r and not C_l by virtue of the fact that there is a continuous path that the right paint ball follows coming from the right machine all the way until it splatters on the canvas. But C_l is culpable$_2$ for E, intuitively because it significantly raised the probability of the fine-grained effect that happened to occur.

14.5.3 LATE CUTTING PREËMPTION

Late cutting preëmption is a special case of preëmption where the preëmption (or fizzling) is the occurrence of the effect. This kind of preëmption is illustrated by replacing the canvas in the previous example with a fragile window and adjusting the machines so that the paint balls are launched with random speeds. Suppose both shots are on target and that the ball from the machine on the right arrives at the window first, shattering it at time t. C_l is culpable$_2$ for E because the firing of the machine on the left significantly raised the probability of (a slightly coarse-grained version of) the actual effect, e. However, we instinctively judge that C_l is not culpable for E because when the window broke, the left paint ball had not yet reached the window. That event counts as a fizzling of C_l's process.

14.6 Culpability$_3$

Culpability$_2$ is deficient because it ignores everything that happens after the candidate cause and before the effect, which makes it unable to take into account the presence of fizzles. However, we do not need to modify the definition of culpability$_2$ much in order to take into account events that happen at other times. To construct a superior concept, culpability$_3$, we simply enlarge the region R in the definition of culpability$_2$ to include what happens at other times, including events located between the candidate cause and effect as well as events occuring after the effect. Often, the additional information acquired includes fizzles that allow us to rule out certain candidate causes.

A definition of culpability$_3$ can now be stated:

> An actual event c (as C qua $\tilde{C}$) is ***culpable$_3$*** for an actual event e (as E) iff a region R (including and surrounding the process leading from c to e) has a contrastive effect $\tilde{E}$ imposed on it that significantly promotes E and includes no fizzling of this process.

This definition differs from the definition of culpability$_2$ primarily by (1) enlarging the region of consideration, R, to include the whole process from c to e and its environment, not just the time of the effect; and (2) forbidding the process heading toward E from fizzling. Presumably, the procedure for evaluating what contrastive event, $\tilde{E}$, is imposed on R needs to be made more sophisticated as well.

The three examples in the previous section included an event c that was judged culpable$_2$ for E but where its process leading to e fizzled. Such events cannot be culpable$_3$ for E because the definition of culpability$_3$ requires the non-existence of the fizzles that were previously cited. So, these examples provide evidence that culpability$_3$ extends culpability$_2$ to accommodate intuitions about causal mechanisms and continuous processes.

It benefits us to have intuitions that match culpability$_3$ rather than culpability$_2$ because culpability$_3$ is not appreciably harder to assess and because it provides more accurate information about prob-influence relations. As illustrated in the examples of preëmption, our intuitions about culpability are likely being driven by perceptions of the paths of the projectiles, so that we are tacitly employing the kind of information captured in culpability$_3$. Imagine we are trying to evaluate the accuracy of the left and right machines in conditions where they are aimed at the same target from very nearly the same location. If we were to try to evaluate their accuracies using culpability$_2$, we would fail because we would not be able to sort out which of the two splotches of paint came from which machine. By assessing culpability$_3$, which one can discern merely by observing the paths of the balls on repeated trials, the accuracy of each machine is equal to the fraction of the trials in which its ball strikes the canvas.

14.7 Culpability$_4$

Culpability$_3$ takes into account intermediate events that *rule out* certain candidate causes as unsuccessful inducers of the effect, but it is also reasonable for humans to have a notion of culpability that takes into account intermediate events that *rule in* additional candidate causes that would not otherwise count as culpable. This more inclusive notion is culpability$_4$. According to the toy theory, culpability$_4$ exists by virtue of an appropriately linked chain of causes. We can define it formally as follows:

> An actual event c (as C qua $\tilde{C}$) is ***culpable$_4$*** for an actual event e (as E) iff there is a chain of culpability$_3$ relations running from c to e.

Because there are no hard and fast rules about which intermediate events count as salient inducers or how they are to be rendered as contrastive events, culpability$_4$ is sensitive to our choices of how to abstract away from the fundamental material layout. By being extremely permissive about event salience, one can achieve an extremely long chain of very finely grained events that are only slightly apart in time. Being extremely permissive as a general policy would result in so many culpability$_4$ relations that culpability$_4$ would have little utility. So, our employment of culpability$_4$ needs to be restricted to a suitably limited class of salient events if we want it to do interesting psychological or explanatory work.

The culpability$_4$ notion should not be thought of as the toy theory's replacement for culpability$_3$ or a decisive improvement on culpability$_3$. Sometimes culpability$_3$ matches our intuitive conception of culpability better than culpability$_4$ and sometimes culpability$_4$ matches it better. The definition of culpability$_4$ ensures that whenever C is culpable$_3$ for E it is also culpable$_4$ for E, but there are often cases where C is culpable$_4$ without being culpable$_3$. These include cases that are widely recognized as counterexamples to the transitivity of causation. In §4.9, I described two scenarios where our intuitions match culpability$_3$ rather than culpability$_4$. But let us consider a simpler example here, adapted from Hall (20):

> A train is rolling along a track that bifurcates and then rejoins after one hundred meters. Suppose that all the relevant details about the background environment are the same on the left side of the track as they are on the right. As the train approaches the junction, the engineer flips a switch that makes the train take the left track. Then, after the train passes the section where the left track rejoins with the right track, the train crosses a road.

Let e (as E) be the event of the train crossing the road, and let c (as C qua $\tilde{C}$) be the activation of the switch for the left track rather than the right track. C is culpable$_4$ for E because $\tilde{C}$ significantly promotes the train moving along the left track, which in turn significantly promotes the train crossing the road, E. Note that the intermediate event does not use the contrast that is fixed by $\tilde{C}$ but is chosen by reckoning salient contrasts at the intermediate time. The switching event is arguably not culpable$_3$ for E because $\tilde{C}$ does not promote the slightly coarse-grained version of e. After all, the probability of the train's reaching the road is the same whether the switch is thrown or not. As with most judgments of culpability$_3$, it is possible in principle to argue that there is some just barely coarse-grained version, E', of the fine-grained effect, e, such that the switching, qua $\tilde{C}$, promotes E' and is thus culpable$_3$ for E. Such an E', though, would need to include the kind of fine details about the likely character of the train had it taken the left track versus the right. For example, there might be more flies near the left track, so that a train taking that route would tend to displace more flies. If such a finely grained construal of the effect were to be countenanced as part of our standards for judging culpability, there would be many more culpable causes than we actually judge. Thus, we can set aside (as too deviant) such a finely grained construal of the effect.

People who are fully aware of what happened in this scenario will be likely to say that the switching event was not one of the causes of E, largely because it is clear that the switching makes no significant difference to how E comes about. Their judgments match what is culpable$_3$. However, if the example is altered slightly, people's judgments will likely match what is culpable$_4$:

> A train is rolling along a track that bifurcates and then rejoins after one hundred meters. As the train approaches the junction, the engineer flips a switch that makes the train take the left track. Then, just after the train starts along the left track, a rare chancy event occurs: a tree standing between the two tracks topples. Given that the tree falls, it has a fifty percent chance of falling across the right track and a fifty percent chance of falling across the left track. The tree happens to fall across the right track, blocking any possible train traffic there. The train crosses the road with no trouble because the train is traveling on the left track.

Our intuitive judgment in such cases is that the switching event was one of the causes of the train successfully crossing the road. Again, the switching is culpable$_3$ for the train's traveling along the left track, which is later culpable$_3$ for the train's making it to the road crossing, E; thus, it is culpable$_4$ for E. And again, the switching would not be culpable$_3$ for E because at the time the switch is thrown, the chance of the train's eventually reaching the road crossing is the same whether it goes along the left track or the right track. Because the switching event does not prob-influence anything concerning the tree, the assessment of what $\tilde{C}$ successfully promoted is not supposed to change (according to the simplistic method for constructing contrastive effects discussed in §14.4). This pair of examples shows that sometimes our common-sense judgments of culpability match culpability$_3$ but not culpability$_4$ and that sometimes they match culpability$_4$ but not culpability$_3$.[36]

When people ask, "What are the causes of E?" they usually do not distinguish between these two different kinds of culpability. But once it is apparent that the toy theory posits these two distinct versions of culpability as guides to our (often presumed to be univocal) implicit notion of culpable cause, it follows that the toy theory has a conflict in the technical sense introduced in §1.8. The toy theory tells us that one good rule of thumb for assessing culpability is that an event is culpable for E iff it is culpable$_3$ for

[36] Because this chapter is only sketching a toy theory of the psychology of culpable causation, not every deficiency can be discussed, but I believe interested readers would benefit from exploring how the motivation I have suggested for distinguishing culpability$_4$ from culpability$_3$ by formulating a more sophisticated scheme for constructing the contrastive effects than the one that I assumed when extending the considerations in §14.4 to handle the temporally extended process leading from the cause to the effect. Specifically, the scheme I presented does not take into account that one's chosen contrast in the cause ought to interact with stuff in the background to help generate the proper contrast to use for representing the effect. If one does so, it may be possible to render the switching event culpable$_3$ for the train's making it to the road crossing, though I suspect people do not reason very clearly about sequences of merely hypothetical interactions beyond simple cases.

E. It also tells us that another good rule of thumb for assessing culpability is that an event is culpable for E iff it is culpable$_4$ for E. Because there are realistic circumstances where an event is culpable$_4$ without being culpable$_3$, the theory provides a conflicting account of which events are culpable. Furthermore, nothing in the toy theory ameliorates this conflict by specifying conditions that adjudicate which version of culpability should supersede the other. According to empirical analysis, these genuine conflicts do not imply that the toy theory is incoherent, nor do they imply that one of the two rules of thumb needs to be rejected as fatally flawed. On the contrary, both versions of culpability have limitations as guides to our cognition of culpable causation, and each offers different benefits. When investigating the psychology of culpable causation at a fairly high level of abstraction, as the toy theory does, it is acceptable to employ RELAXED standards where these kinds of conflicts do not need to be ameliorated with an explicit rule. Insofar as we are just sketching the outlines of a full psychological account, we do not need to specify in every possible instance whether culpability$_3$ or culpability$_4$ is the "correct" account of culpability. And insofar as we are investigating the metaphysics of causation, we do not need an account of culpability at all. The conflict in the toy theory that exists by virtue of its not privileging culpability$_3$ over culpability$_4$ or vice versa does not count as a reason to reject the toy theory qua toy theory.

That point having been noted, nothing in empirical analysis forbids a special science theory from being conflict-free, nor does it discourage our favoring one theory over its rivals for being conflict-free, nor does it countenance scientists against seeking theories that meet STRICT standards of adequacy.

I have already discussed why it is reasonable for people to have intuitions about culpability that match culpability$_3$. Now I would like to cite a few reasons why it is reasonable for people to have intuitions about culpability that match culpability$_4$. For one, fixing plausibly obeys unidirectional transitivity and continuity, as discussed in §4.9 and §4.10, and to a great extent, relations of culpability serve as cognitive proxies for promotion relations. So, it is often convenient for us to think of culpability relations as being continuous and transitive just like the promotion relations they approximate. As Example 14.7 and the two examples from §4.9 demonstrate, it is not correct to think of our intuitive conception of culpability as transitive, but in a wide range of situations, it is convenient to think of C as successfully inducing E by virtue of successfully inducing an intermediate event, which successfully induces another intermediate event, and so on until E occurs. Because the metaphysics of promotion among contrastive events is too complicated for people to manage cognitively, it is understandable that people approximate the unidirectional transitivity of promotion by largely ignoring the background conditions and just thinking of causation as occurring by virtue of a localized chain of events or a localized continuous process.

For a second reason, thinking of culpability as existing by virtue of chains of culpable causation is useful in assembling the full set of events relevant to a causal explanation of an effect that arises through a complicated nexus of events. When there is a sizable set of salient events that play some role in the occurrence of an effect E and we are interested in providing a detailed account of why E occurred, we often not only want to know what events, C_i, were successful inducers of E but also a further explanation of why these C_i occurred, which often involves identifying and citing the events that successfully induced them, and then at a deeper level of explanation the events that successfully induced them. The totality of all such events are the ones that are culpable$_4$ for E. They count as causes of E in the sense of being events that played a significant role in how the total historical development brought about E.

For a third reason, thinking of culpability as existing by virtue of chains of culpable causation serves as a tool in learning about promotion. I will mention just two examples. First, in Example 14.7, the switching event does not promote the train's crossing the road because the chance of the train reaching the road is the same regardless of which track it takes. However, if we judge counterfactual dependence with hindsight,[37] by contrasting what actually happened with what would have happened had the engineer guided the train down the right track, holding fixed the contingency that the tree fell across the right track, then the train's success should count as having counterfactually depended on the engineer directing the train to the left track. In the particular circumstances of this example, such counterfactual reasoning is a misleading guide to the promotion relations because the switching event did not improve the chances that the train would make it safely to the road. However, in a wide variety of cases, after-the-fact events such as the tree falling are indicative of the existence of some hidden condition of earlier states. When a tree falls toward the right in a seemingly spontaneous manner, that is often because there is some hard-to-identify-in-detail feature of the previous condition of the tree that induces its falling at that time to the right. If the tree's falling to the right were due to such a condition rather than brute chance, it would be correct to say the switching event successfully promoted the train's crossing the road. So, because we instinctively judge counterfactual dependence by presuming that the tree would still have fallen across the right track if the train had gone to the right, we often succeed at inferring the correct promotion relations, promotion relations that we would never be able to detect if we restricted our attention to what was happening at the time the switching event occurred.

For a second example of how culpability$_4$ serves as a heuristic for learning about promotion, consider causation that occurs via some enabling (or disabling) condition. Promotion stemming from enabling or disabling conditions are sometimes difficult to detect, and culpability$_4$ helps us filter through possible candidates more quickly. An enabling condition can be thought of as an event that is normally considered part of the background and induces an effect E in the presence of a more salient inducer of E, which counts as an activating condition. A disabling condition is similarly an inhibitor that lies in the background. For example, we might recognize that some migratory species, say the canvasback duck, annually visits a certain lake for mating. One year, the ducks do not successfully reproduce. That should lead us to suspect that there is some inhibitor of duck reproduction, perhaps in the water. Because there are many chemicals in the water, it might be difficult to identify what, if anything, inhibited the reproduction. However, if we can see that a factory is pouring some sort of liquid into

[37] For similar observations, see Edgington (16), Kvart (39), Northcott (52), and my discussion of "infection by culpable causation" in the supplementary material I have provided concerning Morgenbesser's coin.

the lake, then it is reasonable to suspect that a chemical from the factory is culpable$_3$ for the condition of the water. Because we have previously learned that waterborne chemicals are sometimes culpable$_3$ for reduced bird reproduction and because we know just by looking that the factory is plausibly culpable$_3$ for some sort of effect on the watershed, we are justified in inferring that there is a reasonable possibility that the factory is culpable$_4$ for the failure of the ducks to reproduce. This can justify restricting the testing to chemicals used in the factory instead of testing for the full array of epistemically possible chemicals in the lake. If it were illegitimate to identify potential causes by using what we know about chains of culpability, we might waste time testing other possible sources. For example, if we can be sure that the chemicals stored in some nearby warehouse never left the warehouse, we can be reasonably sure that they are not culpable for any effects on the water supply, and thus reasonably sure that they are not culpable for the canvasbacks' troubles. This indicates that it is likely unnecessary to test the water for these chemicals.

One of the consequences of having both culpability$_3$ and culpability$_4$ is that many questions about culpability that might initially seem straightforward become extremely messy. An exemplary complicated case is Hesslow's (29) thrombosis example. Taking a birth control pill regularly is a promoter of thrombosis by virtue of its direct role as chemical in the body. But the birth control pill is also an inhibitor of pregnancy, which itself is a promoter of thrombosis. So, there are two routes by which thrombosis is probabilistically influenced. For the sake of discussion, we can modify the example to have them approximately cancel each other out over the course of time, so that taking the pills on the whole has no net probabilistic influence on thrombosis. Imagine that some woman takes the birth control pill, does not become pregnant, and is not afflicted by thrombosis. Is her consumption of the pills one of the causes of her being free of thrombosis? It might seem that the pills cannot be culpable$_3$ for her failure to contract thrombosis because they do not prob-influence thrombosis. It might also seem that the pills are culpable$_4$ for her not getting thrombosis because there are many chains of successful promotion that run from her taking a given pill to her lack of thrombosis at later times. So, the two notions conflict in their attribution of causal culpability. A univocal assessment is made even more difficult when we take into account that taking pills for a full year consists of many localized events: the daily occurrences where she ingests a single pill. It is plausible that many of these events exert different degrees of promotion and inhibition through different intermediate mundane events. Furthermore, whether an event is culpable$_4$ depends on which events are permissible for employment in chains of culpable$_3$ causation. Remember that if we identify salient events liberally, allowing all sorts of non-standard contrasts and coarse-grainings, just about any event will count as culpable$_4$ for her not having thrombosis. So, the relevant culpability$_4$ would have to be restricted to some appropriately salient events in order to match our psychological judgment that there do not exist a vast multitude of thrombosis preventers. But we do not have clear intuitions about how to break down vast causal networks (like those present in the daily operation of the human body) into relevant component events. In summary, it is safe to say that in complicated interactions like those exhibited by the thrombosis example, it is difficult to make unequivocal statements about culpable causation that are well grounded in our practices of attributing culpability.

At this point, my presentation of the toy theory is complete. There are several more features of the toy theory that constitute additional evidence that the toy theory meshes well with my account of the metaphysics. The degree of support provided by each of the following sections to the metaphysical theory of causation is individually small, but I think each discussed feature helps to reinforce my contention that the toy theory of causal culpability is not concocted ad hoc to accommodate psychological data but is reasonably well motivated by the hypothesis that we have a conception of culpable cause because it provides a useful shortcut for learning about promotion. The following sections are meant as stand-alone commentaries and are not presented in any special order.

14.8 Uncaused Events

'Uncaused events' can be understood as events with no culpable causes.[38] Uncaused events can occur when an event sensitively depends on many parts of a previous state without there being a promoter that fits a relatively simple natural language description. Uncaused events can inhabit both deterministic and indeterministic universes. One example is when a fuse spontaneously lights due to a fantastically unlikely thermal fluctuation. Another example is when an evaporating salt solution forms a crystal that lines up in some direction. The alignment of the crystal is uncaused in the sense that its direction results from the chance arrangement of some initial conjoining atoms, followed by other atoms aligning with the seed crystal.

One nuance that deserves brief mention is the case of magnetic resonance imaging (MRI) described in §5.6. The event where the glycerin sample emanates electromagnetic radiation is preceded by a state, s, occurring at the time the 'FLIP' signal is broadcast. In the region of the glycerin sample, the state s is macroscopically unremarkable. Due to the extremely intricate pattern of particle spin directions, even a very slight contextualization of s is very unlikely to promote an emanation of radiation from the sample. Superficially, it seems that the passive return to a nearly perfectly aligned pattern of particle spin axes should count as a remarkable coincidence like an anti-thermodynamic fluctuation and hence should count as uncaused. However, there is a previous alignment event that leads to the intermediate state s in a way that can be reliably affected by the 'FLIP' signal to emanate radiation from the sample. Because of the practical reliability of the MRI causal process, we can count s as falling under a relatively natural language description by courtesy. Unlike spontaneous anti-thermodynamic fluctuations, it is involved in reliable promotion regularities. The point here is merely that there is a bit of subtlety in what counts as a culpably uncaused event. Furthermore, this subtlety bears on the definition of fizzling. Under normal conditions for judging whether a process has fizzled, states like s would count decisively in favor of the MRI process having fizzled because anything more than the very slightest coarse-graining (with contextualization) of s would fix a low probability for

[38] Remember that no events are altogether uncaused. Every fine-grained event trivially termines itself, and every contrastive event trivially promotes itself.

ECHO. However, it makes sense to ignore states like *s* when judging fizzling because these kind of states are reliably produced by ALIGN and reliably result in ECHO. The unusual character of *s* is in practice inaccessible by observation of *s* alone, but because we have so many successful cases of promotion that run from ALIGN to FLIP to ECHO, it is clear that the development from FLIP to ECHO is no mere accident. Thus, we ought to incorporate into our rules for fizzling an exception clause to handle special cases like *s*, where the intermediate structures needed for promotion are present but are hidden in the microstate.

14.9 Prevention

A characterization of prevention that I believe is adequate for the purposes of the toy theory is as follows: An event C prevents E iff C occurs, E does not occur, and C is a salient significant inhibitor of E. The expression 'salient significant inhibitor' is supposed to be understood as qualified in §14.2.

The factors that affect salience for prevention are much like those for ordinary cases of culpable causation. For example, consider some scenarios tested by Walsh and Sloman (67):

> There is a bottle at the bottom of a hill. Frank is standing close by at the top. While he is there, Billy aims to roll a ball toward the bottle. The aim is perfectly on target. Billy lets go of the ball and it rolls down toward the bottle. Frank then runs down the hill after the ball. He manages to catch up with the ball and picks it up before it reaches the bottle. The bottle does not break. Did Frank prevent the bottle from breaking?

> There is a bottle at the bottom of a hill. Billy is standing close by at the top. While he is there he thinks about rolling a ball toward the bottle. He always has a perfect aim and he will definitely hit the bottle. At the last minute Billy changes his mind. He decides not to roll the ball. The bottle does not break. Did Billy prevent the bottle from breaking?

> There is a bottle at the bottom of a hill. John is standing close by at the top. While he is there, John aims to roll a ball toward the bottle. The aim is perfectly on target. John lets go of the ball and it rolls down toward the bottle. Within a split second he then chooses to run down the hill after the ball. He manages to catch up with the ball and picks it up before it reaches the bottle. The bottle does not break. Did John prevent the bottle from breaking?

The subjects of the experiments answered the questions affirmatively at rates of 84%, 70%, and 46%, respectively. According to the toy theory, it is understandable that people tend to believe that Frank prevented the bottle from breaking. There was a salient event, his picking up the ball, and it successfully inhibited the breakage. It is somewhat understandable that Billy was less often judged to have prevented the bottle from breaking. That is presumably because people do not conceive of Billy's possible roll as something that he ought to be doing, so his decision not to roll the ball counts as not salient. It is also understandable that people tend to believe John prevented the bottle from breaking. There was a salient event, which was his picking up the ball, and it successfully promoted the lack of breakage. The subjects' reduced positive response in the third scenario can be explained by the question's multiplicity of potential preventers: rolling the ball, picking it up, or both. Also, some subjects might have focused on John's picking the ball up as the potential preventer but then imported a contrast where John did not do anything at all rather than a contrast where John rolled the ball but did not pick it up, which would count John as not having prevented the breakage.

Regarding potential improvements on the toy theory's definition of prevention, it is clear that some kinds of preventions take place via some recognizably continuous process. In those cases, the process can fizzle, leading to a judgment of no prevention. To what extent the default rules for fizzling hold for cases of prevention is a topic beyond the scope of this discussion.

14.10 Double Prevention

Double prevention occurs when some C prevents an intermediate event I from occurring that would have prevented E, had I occurred. Causation by double prevention occurs when C is a culpable cause of E by virtue of double prevention issuing from C to E.

The only new conceptual resource needed for understanding double prevention is that we need to make sense of counterfactual prevention. The intermediate event is an event that *would have* prevented E. To make sense of claims about such counterfactual prevention, we can first make explicit that C is contrastivized as $\tilde{C} \equiv (\overline{C}, \overline{\neg C})$. Then, we can model the non-occurrence of I as the contrastive event $\tilde{N} \equiv (\overline{\neg I}, \overline{I})$, where $\overline{\neg I}$ is the contextualized event fixed by $\overline{C}$ and $\overline{I}$ is a contextualization of the actual state of the world at the time of $\overline{\neg I}$. The promotional link needed for double prevention is for $\tilde{C}$ to be a salient promoter of $\tilde{N}$ and for $\tilde{N}$ to be a salient promoter of E. The other necessary conditions for double prevention to count as culpable causation are that C and E occur, and that I not occur. Interestingly, there is a good case to be made that normally $\tilde{N}$ will count as a salient with regard to its promotion of E. That is because its contrast is the actual state of the world. It is reasonable to think that when we are evaluating non-actual events for what they promote, the actual state of the world is automatically a salient contrast.

Some cases of double prevention strike people as clear instances of causation. For example, as discussed in Schaffer (57), a gun trigger is constructed so that C, pulling the trigger, prevents I, the gun's sear from being located in its normal place. The event I would prevent a spring from uncoiling and causing the explosion that propels the bullet. Suppose the trigger is pulled, and the gun fires. In this case, people will readily claim that pulling the trigger caused the gun to fire. This agrees with the toy theory's assessment that C successfully promoted $\neg I$, which successfully promoted E.

Some instances of double prevention tend to strike people as not being clear cases of causation. For example, Steve removes the sign labeled 'Danger' from the beach. Then, later in the day, Mark goes surfing. Mark would not have surfed if he had seen the danger sign. So C, Steve's removal of the sign prevented I, the sign's being in its default location, and I in turn would have prevented E, Mark's surfing. Is Steve's action a cause of Mark's surfing? People will not normally judge that Steve caused Mark to surf. However, people might have a tendency to think that Steve's action was one of the causes of Mark's surfing. The explanation of this phenomenon is that the example plays off of the ambiguity of the phrase 'is a cause of', which I earlier noted is ambiguous between 'is one of the causes of' and 'is something that caused'. Steve's action can count as a culpable cause of Mark's surfing but does not cause Mark to surf because it is not a prominent promoter of Mark's surfing.

In still other cases, double prevention is clearly not a case of culpable causation. Suppose Dave is knowledgeable about dangerous ocean currents, and would have prevented Mark from surfing if he had been at the beach. But, as it happens, Dave was working at his shop as he usually does. Dave's working prevented him from going to the beach where he would have prevented Mark from surfing. People will not cite Dave as one of the causes of Mark's surfing, but this results from a lack of salience. Dave's decision to be at work is not culpable for his being absent from the beach because it is not a salient event; he is doing what he is supposed to be doing. If Dave were obligated to serve as a lifeguard at the beach, then we could reasonably cite his presence at the shop as one of the causes of Mark's surfing.

14.11 Culpable Causation by Omission

An omission is something that does not happen, so it might appear to have no role in causation. Yet, omissions are routinely cited throughout ordinary language and in science as causes, and they play largely the same roles as ordinary causes in prediction, manipulation, promotion, explanation, and culpability. The philosophical literature based on orthodox analyses of causation often cites causation by omission as a potential problem case, and my comments in this section are aimed at showing how causation by omission is unproblematic.

For purposes of discussion, any cause that is paradigmatically not an omission may be said to be a ***positive cause***. Examples of omissions include a lack air and the absence of the financial officer at the board meeting. Positive causes include an abundance of air and the presence of the financial officer at the board meeting.

According to my account, omissions are metaphysically exactly like positive causes.

Both are instantiated by some fundamental material contents somewhere in the arena. The only difference between omissions and positive causes is descriptive. Describing a fine-grained event as a positive event communicates that we intend to construe it as a contrastive event with a default contrast, a contrast that instantiates a contextually relevant absence of the positive event. Describing a fine-grained event as an omission, however, typically indicates an interest in a non-default contrast. (See the earlier discussion in §4.7 of the role omissions play in promotion. Also see Schaffer (60) for a similar account.)

I will now describe three examples to illustrate how omissions can be culpable causes. First, imagine the following scenario illustrating a positive cause. There is a hungry tiger in a child's bedroom and the tiger sees the child and starts licking his foot. In this case, it is easy to see how the following claim of culpable causation is true.

> (1) The presence of a hungry tiger in a child's bedroom caused the sleeping child to wake up.

For contrast, consider a case of causation by omission. In this second example, a zookeeper has just been told that someone saw a tiger in the main elephant pen so she examines the main elephant pen and finds the two bull elephants but no tiger. It turns out that the zoo's tigers are in their proper pen. Consider the following.

> (2) The lack of a hungry tiger in the main elephant pen caused the zookeeper to feel relieved.

A natural contrastivization $\tilde{C}$ of the lack of a hungry tiger is formed by contextualizing the actual state somehow and constructing a contrast contextualized event that is different by having the tiger moved from wherever it actually is to the main elephant pen, filling in the environmental details in a natural way. It follows from the fundamental laws and a reasonable interpretation of the scenario that $\tilde{C}$ successfully promoted the relief of the zookeeper.

Notice that so far, the absence of a tiger in the second example was treated exactly parallel to the positive presence of the tiger in the first example. The only difference was in the pragmatics of how the contrast state was picked out. In the first example, the tiger is present, so to form the contrast we subtract the tiger from the rest of the environment. In the second example, there are two elephants present in the pen, so to form the contrast we just shift the position of an actual tiger from its pen into the elephant pen. If someone had walked up to the pen without having been told the false story about the misplaced tigers and had been asked what kinds of things are interacting in a 'cause and effect' relation, they would not think to identify the event 'absence of a hungry tiger' as a cause of anything. They would cite the elephants, and maybe the straw or watering bin. They would just use their standard heuristics for decomposing a scene into objects and their claims about causation would implicitly employ contrasts formed by just striking out those objects in a natural way. By explicitly describing the elephant pen as instantiating an absence of the tiger, the normal ways we pick out contrasts is overridden in order to evoke the presence of the tiger as a conversationally suggested contrast.

So what is the problem with causation by omission? To see this, let us look at a third case, where a child has been playing all day long and is much more tired than normal and is now in his bedroom. There is no tiger anywhere remotely nearby, and the child sleeps soundly throughout the night unlike most nights where the child intermittently wakes. Ordinary people, if given as much information as desired about the circumstances and asked to identify what caused the child to sleep soundly, would cite that it was night and that the child played more than usual and other positive causes. They would not find the following claim agreeable:

> (3) The lack of a hungry tiger in the child's bedroom was one of the causes of the child sleeping soundly.

According to orthodox standards, theories of causation are supposed to be judged on their ability to reproduce folk assessments of singular causal claims in clear cases. And prima facie, this is a clear case. So, at first blush, a theory of causation is required to assess this statement as explicitly false. The problem for orthodox theories is that it is often difficult to find a theory for how the absence of the hungry tiger can count as a cause in (1), but not count as a cause in (3). They seem to embody many of the same features.

The explanation provided by my toy theory is straightforward. The fine-grained event c, which instantiates the lack of a tiger, does contribute to the child's sleeping. The contrastive event $\tilde{C}$, which represents the absence of a tiger rather than the presence of a tiger, does promote the child's sleeping. But C does not count as culpable because $\tilde{C}$'s contrast is not salient: there is no reason to consider the possible presence of tigers. Thus, the lack of a hungry tiger does not count as a culpable cause of the child's sleeping soundly.

To summarize, the role of omissions can be subdivided in terms of the three conceptual layers of causation. (1) Uncontroversially, the fundamental events that contributed to the child's sleeping did not instantiate a hungry tiger in the room. (2) Uncontroversially, the absence of a hungry tiger in a child's bedroom does raise the probability that the child will sleep soundly. (3) Assuming the presence of a tiger is not contextually relevant, citing the absence of a hungry tiger as a culpable cause is as inappropriate as citing the absence of an alligator, the absence of a marching band, the absence of a helicopter crash, etc. In order to avoid bogging down conversation by citing an infinite list of all the absences that promoted the effect, we have a convention whereby events that are not contextually salient are set aside. That suffices for an adequate account. Having explained how the omission works with respect to the three layers of causation, the explanation of causation by omission is complete.

14.12 Summary

The toy theory of our psychology of culpable causation that I presented in this chapter was an attempt to connect our folk intuitions about causal culpability to the metaphysics of causation, especially promotion. The toy theory is deliberately sketchy and vague in numerous respects and to the extent that it is precise enough to make predictions, it is surely in conflict with some psychological data. My goal was merely to provide an example of how to approach an empirical analysis of the non-metaphysical aspects of causation. Such an analysis should not try merely to systematize people's judgments about what events are singular causes or are relevant to causal explanations of singular effects but should try to connect this data to the metaphysics of causation. The pair of empirical analyses together ought to make sense of how our intuitions and reasoning about causation help us track metaphysical relations like promotion.

{ REFERENCES }

Adams, E. (1975). *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: D. Reidel. 5

Adams, E. (1976). "Prior Probabilities and Counterfactual Conditionals," in W. L. Harper and C. A. Hooker (eds.), *Foundations of Probability Theory, Statistical Inference and Statistical Theories of Science* 1, 1–21. 5

Adams, E. (1998). *A Primer of Probability Logic*. Stanford: CSLI Publications. 5

Alicke, M. (1992). "Culpable Causation," *Journal of Personality and Social Psychology* 63 (3), 368–378. 42, 62

Barker, S. (1999). "Counterfactuals, Probabilistic Counterfactuals, and Causation," *Mind* 108, 431–469. 18

Bennett, J. (1984). "Counterfactuals and Temporal Direction," *The Philosophical Review* 93 (1), 57–91. 18

Chalmers, D. and Jackson, F. (2001). "Conceptual Analysis and Reductive Explanation," *The Philosophical Review* 110 (3), 315–360. 36

Collins, J. (2000). "Preemptive Prevention," *Journal of Philosophy* 97, 223–234, reprinted in (2004) J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press. 38

Collins, J.; Hall, N.; and Paul L. A., eds. (2004). *Causation and Counterfactuals*. Cambridge: MIT Press. 37, 45, 46, 52

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press. 43

Dowe, P. and Noordhof, P. (2004). *Cause and Chance: Causation in an Indeterministic World*. London: Routledge.

Driver, J. (2008a). "Attributions of Causation and Moral Responsibility," in W. Sinnott-Armstrong (Ed.) *Moral Psychology (Vol. 2): The Cognitive Science of Morality: Intuition and Diversity* 423–439. Cambridge, MA: MIT Press. 62

Driver, J. (2008b). "Kinds of Norms and Legal Causation: Reply to Knobe and Fraser and Deigh," in W. Sinnott-Armstrong (ed.), *Moral Psychology (Vol. 2): The Cognitive Science of Morality: Intuition and Diversity* 459–461). Cambridge, MA: MIT Press. 62

Ducasse, C. (1926). "On the Nature and the Observability of the Causal Relation," *Journal of Philosophy* 23 (3), 57–68. 35

Edgington, D. (1995). "On Conditionals," *Mind* 104 (414), 235–329. 5

Edgington, D. (2004). "Counterfactuals and the Benefit of Hindsight," in P. Dowe and P. Noordhof (eds.), *Cause and Chance: Causation in an Indeterministic World*. London: Routledge. 18, 28, 83

Elga, A. "Statistical Mechanics and the Asymmetry of Counterfactual Dependence," *Philosophy of Science* 68 (3) Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers, S313–S324.

Gauker, C. (2005). *Conditionals in Context*. Cambridge, MA: MIT Press. 3, 6

Goodman, N. (1947). "The Problem of Counterfactual Conditionals," *Journal of Philosophy* 44, 113–128. 6

Hall, N. (2000). "Causation and the Price of Transitivity," *Journal of Philosophy* 97, 198–222. 38, 80

Hall, N. (2002). "Non-locality on the Cheap? A New Problem for Counterfactual Analyses of Causation," *Noûs* 36, 276–294. 38

Hall, N. (2004). "Two Concepts of Causation," in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press. 37, 38, 42

Hall, N. (2007). "Structural Equations and Causation," *Philosophical Studies* 132, 109–136. 62

Halpern, J. Y. and Pearl, J. (2005). "Causes and Explanations: A Structural-model Approach. Part I: Causes," *The British Journal for the Philosophy of Science* 56, 843–887. 43

Halpern, J. Y. and Pearl, J. (2005). "Causes and Explanations: A Structural-model Approach. Part II: Explanations," *The British Journal for the Philosophy of Science* 56, 889–911. 43

Harper, W. L., Stalnaker, R., and Pearce, G., eds. (1981). *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Dordrecht: D. Reidel.

Hart, H. L. A. and Honoré, T. (1959). *Causation in the Law*. Oxford: Clarendon Press. 42

Hausman, D. (1998). *Causal Asymmetries*, Cambridge: Cambridge University Press. 43

Hesslow, G. (1981). "Causality and Determinism," *Philosophy of Science* 48, 591–605. 84

Hitchcock, C. (2001). "The Intransitivity of Causation Revealed in Equations and Graphs." *Journal of Philosophy* 98, 273–299. 43

Hitchcock, C. (2003). "Of Humean Bondage," *The British Journal for the Philosophy of Science* 54, 1–25. 54

Hitchcock, C. (2007). "Three Concepts of Causation," *Philosophy Compass* 2/3: 508–516. 42

Hitchcock, C. (2009). "Structural Equations and Causation: Six Counterexamples," *Philosophical Studies* 144, 391–401. 62

Huemer, M. and Kovitz, B. (2003). "Causation as Simultaneous and Continuous," *The Philosophical Quarterly* 53 (213), 556–565. 65

Jackson, F. (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press. 36, 38

Knobe, J. and Fraser, B. (2008). "Causal Judgment and Moral Judgment: Two Experiments," in W. Sinnott-Armstrong (ed.), *Moral Psychology (Vol. 2): The Cognitive Science of Morality: Intuition and Diversity* 441–447. Cambridge, MA: MIT Press. 62

Kvart, I. (1986). *A Theory of Counterfactuals*. Indianapolis: Hackett Publishing. 27

Kvart, I. (1994). "Causal Independence. *Philosophy of Science* 61, 96–114. 27

Kvart, I. (2004). "Probabilistic Cause, Edge Conditions, Late Preemption and Discrete Cases," in P. Dowe and P. Noordhof (eds.), *Cause and Chance: Causation in an Indeterministic World*. London: Routledge. 83

Lewis, D. (1973a) *Counterfactuals*. Oxford: Blackwell. 3, 6, 7, 8, 20

Lewis, D. (1973b). "Causation," *Journal of Philosophy* 70 556–567, reprinted in (1986) *Philosophical Papers, Volume 2*. Oxford: Oxford University Press. 20, 21, 37, 43

Lewis, D. (1979). "Counterfactual Dependence and Time's Arrow," *Noûs* 13 455–476, reprinted in (1986) *Philosophical Papers, Volume 2*. Oxford: Oxford University Press. 6, 8, 18, 20, 21

Lewis, D. (1986). "Postscripts to 'Causation,'" *Philosophical Papers, Vol. II* Oxford: Oxford University Press. 36

Lewis, D. (2004). "Causation as Influence," as revised in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press. 36, 43, 46

Loewer, B. (2007). "Counterfactuals and the Second Law," in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press. 24

Mackie, J. L. (1973). *The Cement of the Universe*. Oxford: Oxford University Press. 43

Maudlin, T. (2004). "Causation, Counterfactuals, and the Third Factor," in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press. Reprinted in (2007) *The Metaphysics in Physics*. Oxford: Oxford University Press. 62

Maudlin, T. (2007). *The Metaphysics in Physics*. Oxford: Oxford University Press. 18, 26, 27

Menzies, P. (2007). "Causation in Context," in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press. 43

Noordhof, P. (2004). "Prospects for a Counterfactual Theory of Causation," in P. Dowe and P. Noordhof (eds.), *Cause and Chance: Causation in an Indeterministic World*. London: Routledge. 27

Noordhof, P. (2005). "Morgenbesser's Coin, Counterfactuals and Independence," *Analysis* 65, 261. 27, 41

Northcott, R. (2010). "Natural Born Determinists: a New Defense of Causation as Probability-raising," *Philosophical Studies* 150, 1–20. 83

Paul, L. A. (2000). "Aspect Causation," *Journal of Philosophy* 97, 235–256, reprinted in (2004) J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press. 38

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press. 43

Schaffer, J. (2000a). "Overlappings: Probability-Raising without Causation," *Australasian Journal of Philosophy* 78, 40–46. 52, 70, 77

Schaffer, J. (2000b). "Trumping Preemption," *Journal of Philosophy* 97, 165–181, reprinted in (2004) J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press. 71

Schaffer, J. (2000c). "Causation by Disconnection," *Philosophy of Science* 67, 285–300. 87

Schaffer, J. (2001). "Causes as Probability Raisers of Processes," *Journal of Philosophy* 98, 75–92. 60

Schaffer, J. (2004). "Counterfactuals, Causal Independence and Conceptual Circularity," *Analysis* 64 (284), 299–308. 28, 41

Schaffer, J. (2005). "Contrastive Causation," *The Philosophical Review* 114 (3), 327–358. 89

Skyrms, B. (1981). "The Prior Propensity Account of Subjunctive Conditionals," in W. L. Harper, R. Stalnaker, and G. Pearce (eds), *Ifs*. Dordrecht: D. Reidel, 259-65. 5

Sloman, S. (2005). *Causal Models: How People Think about the World and its Alternatives*. Oxford: Oxford University Press. 44

Slote, M. (1978). "Time in Counterfactuals" *The Philosophical Review* 87, 3–27. 41

Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge: MIT Press. 43

Suppes, P. (1970). *A Probabilistic Theory of Causality*. Amsterdam: North-Holland. 43, 49

Talmy, L. (1988). "Force Dynamics in Language and Cognition," *Cognitive Science* **12**, 49–100. 60

Walsh, C. and Sloman, S. (2011). "The Meaning of Cause and Prevent: The Role of Causal Mechanism," *Mind & Language* **26** (1), 21–52. 86

Weslake, B. (2006). "Common Causes and the Direction of Causation," *Minds and Machines* **16** (3), 239–257. 44

Wolff, P. and Zettergren, M. (2002). "A Vector Model of Causal Meaning," *Proceedings of the twenty-fifth annual conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum. 60

Wolff, P. (2007). "Representing Causation," *Journal of Experimental Psychology: General* **136** (1), 82–111. 60

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press. 43, 44

Yablo, S. (1992). "Mental Causation," *The Philosophical Review* **101**, 245–280. 63

Yablo, S. (1997). "Wide Causation," *Philosophical Perspectives: Mind, Causation, and World* **11**, 251–281. 63

{ INDEX }

actuality-focusing, 10
antecedent event, 12

backtracking counterfactuals, 65
bizarreworld, 6, 20, 22

causalasymmetry, 64–65; chain, 61, 79;
 explanation, 83; irreflexivity, 64;
 mechanism, 59; process, 59, 77
causal grouping, 66–69
causationculpable, 59; general, 42, 59;
 non-metaphysical aspects of, 91;
 orthodox metaphysics of, 50; singular,
 42, 59
causation by omission, 88–90
causeculpable, 42
chain, causal, 61, 79
conceptual analysisorthodox, 35
conditionalnomic, 1
conflict, 81
conflict ameliorationdelegation of, 58
consequent event, 12
continuity, 82
counterfactual, 41 mundane, 12
counterfactual dependence, 59
counterlegal, 19
culpability, 50 as proxy for promotion, 57, 59;
 tension in concept of, 59
culpability$_1$, 61, 62, 69, 72
culpability$_2$, 61, 72, 73, 75–78
culpability$_3$, 61, 78–82, 84
culpability$_4$, 61, 79–84
culpable causation, 59
culpable cause, 42

difference-making, 59
double prevention, 87–88

empirical analysisof the non-metaphysical
 aspects of causation, 91
enabling cause, 83
eventmundane, 12, 42
examplebalance scale, 67, 68; bridge collapse,
 63; duck reproduction, 83, 84; fired gun
 as double prevention, 88; fizzled rocket
 fuse, 77; lit panel button, 63; lottery ticket,
 66; match striking, 62; neutrino-triggered
 bomb, 18; paint balls, 70–71, 73–76, 78;
 surfing, 88; thrombosis, 84; tiger, 89–90;
 track switching, 80–81, 83; tree falling on
 track, 81
exemplars, 35
explicitly true, 36, 48

fizzle, 77; saved, 77
fizzling, 85
focal set, 62
force dynamic model of causation, 60

general causation, 42, 59

imposition, 75, 79
inducement, 59, 60, 74; successful, 59, 61
inus account of causation, 43, 47

joint causation, 68–69

managed inconsistency, 58
mechanism, causal, 59
minimal account of counterfactuals, 16
Minkowski space-time, 27
morality and causation, 62
Morgenbesser's coin, 6, 24–31, 33–34, 41
mundane counterfactual, 12
mundane event, 12, 42

new-fangled orthodox analysis, 37
nomic conditional, 1
non-spatiality, 27

old-fashioned orthodox analysis, 35
omission, 88
orthodox (conceptual) analysis, 35
 new-fangled, 37; old-fashioned, 35
overdetermination, 66–68

positive cause, 88
prevention, 86
preëmption, 48, 77–78; early cutting, 78; late
 cutting, 78
probability-raising, 59
production, 59

RELAXED standardsof theoretical adequacy,
 82

salient promoter, 62
saved fizzle, 77

significant promotion, 65–66
simultaneous causation, 65
singular causation, 42, 59
SM-conditional, 24
space-timeMinkowski, 27
sufficiently specified antecedent, 13

thrombosis example, 84
top conceptual layerof causation, 57
transitivity, 82
trumping, 71

underspecified antecedent, 12